Vers une approche semi-automatique pour la définition de motifs d'argumentation utilisés dans les résumés de projets scientifiques du domaine de la biodiversité

Rosa Cetro, Marc M. Barbier, Philippe P. Breucker, Hilde Eggermont, Philippe Gambette, Tita Kyriacopoulou, Xavier Le Roux, Claude Martineau, Nicolas N. Turenne

▶ **To cite this version:**

Rosa Cetro, Marc M. Barbier, Philippe P. Breucker, Hilde Eggermont, Philippe Gambette, et al.. Vers une approche semi-automatique pour la définition de motifs d'argumentation utilisés dans les résumés de projets scientifiques du domaine de la biodiversité. Revue des Nouvelles Technologies de l'Information, 2014, Fouille de Données et Humanités Numériques, RNTI-SHS-2 (2), pp.47-80. hal-01090607v2

HAL Id: hal-01090607
https://hal.science/hal-01090607v2

Submitted on 16 Oct 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Vers une approche semi-automatique pour la définition de motifs d'argumentation utilisés dans les résumés de projets scientifiques du domaine de la biodiversité

Rosa Cetro*, Marc Barbier **, Philippe Breucker **, Hilde Eggermont***,¥, Philippe Gambette*, Tita Kyriacopoulou*, Xavier Le Roux****,¥, Claude Martineau*, Nicolas Turenne**@

* Université Paris Est - UMR LIGM (Laboratoire d'Informatique Gaspard Monge)
** INRA Sens, IFRIS, Université Marne la Vallée (France)
*** Belgian Biodiversity Platform, Belgian Science Policy Office
**** Université de Lyon 1 - UMR Ecologie Microbienne
¥ERAnet Biodiversa
@ contact : nturenne.inra@yahoo.fr

Résumé. Nous positionnons notre travail dans le domaine de l'analyse et de la visualisation de données textuelles produites par les scientifiques et réunies en corpus calibré. Ce domaine est reconnu pour sa contribution à la réflexion sur la composition et l'évaluation des politiques scientifiques. Le corpus que nous utilisons est une collection de tous les résumés de projets acceptés dans des guichets d'appels à projet dans le domaine de la biodiversité référencés par le réseau européen BiodivERsA. L'objectif de ce travail ancré dans la sociologie des sciences consiste à mieux comprendre les principales caractéristiques utilisées par les scientifiques pour présenter leur projet et convaincre de ses qualités. Pour cela nous avons utilisé une pluralité d'outils face à la difficulté de dépouiller l'information pour associer le niveau sémantique (structure de l'information) au niveau pragmatique (relations entre les rédacteurs de projet). Notre contribution repose sur un nouveau type d'extraction d'information, hors entités nommées, basé sur l'extraction de motifs d'argumentation. D'une part on remarque que l'usage de ces motif marque la présence d'arguments dans des résumés de projets, et d'autre part croît avec le temps.

1 Introduction

Les projets de collaboration scientifique ont été très peu étudiés par rapport à la littérature scientifique dans le cadre STS (science and technology studies) /SPS (social and political science). Toutefois, prendre l'entité « projet » comme objet d'étude est particulièrement pertinente à l'heure où l'organisation par projet constitue un modèle de financement et un modèle d'organisation pour la recherche publique. Dans cette contribution nous présentons une proposition d'étude de ce genre textuel, en nous appuyant sur un corpus de projets tirés de la base de données Biodiversa (www.biodiversa.org/8).

Une approche pour analyser un contenu textuel consiste à étudier la distribution des termes clés et d'en faire soit un inventaire soit un regroupement pour établir des tendances thématiques [Lent & Agrawal, 1997] [Turenne & Barbier, 2004] [Mei & Zai, 2005] [Song et Kim, 2012]. Cette problématique de structuration documentaire est intéressante pour con-

naître un contenu hiérarchisé et spécifique à des ensembles de documents. En ce qui nous concerne on veut savoir quelles informations dans un document - projet sont mises en valeur pour qu'il soit considéré comme élément de valeur dans la base projet. A ce titre la citation d'un terme clé thématique peut être déterminante mais c'est une hypothèse restrictive. Notre hypothèse est donc de considérer la structure argumentaire d'un document en procédant à une définition de motifs d'argumentation et en testant si ces motifs d'argumentation (supposés défendre l'intérêt du projet) sont régulièrement employés dans les documents. Cela reprend l'idée de [Searle, 1969], en philosophie, qui énonce le lien fort entre l'énoncé et l'acte d'élocution. Dans notre cas, cela se traduit par une relation entre le statut d'un projet et l'extraction automatique des formes du discours. Si tel est le cas, l'utilisation de motifs d'argumentation s'avère être un moyen structurant décisif pour proposer un projet. De la rhétorique à la philosophie politique en passant par l'étude des énoncés du discours pour l'étude de la communication [Breton et Gauthier, 2000] retracent l'histoire des principales théories de l'argumentation [Perelman et Olbrechts-Tyteca, 1958] [Toulmin, 1959] [Goffman, 1981] [Anscombre et Ducrot, 1983] [Culioli, 1999]. La logique du premier ordre est aussi prolixe en modèles d'argumentation pour l'étude du raisonnement automatique, mais sans considérer des données textuelles et ce qu'on y trouve par exploration [Caminada et Amgoud, 2007] [Besnard et Hunter, 2008] [Toni, 2008] [Ontanon et Plaza, 2010] [Kaci, 2010] [Toniolo et al, 2012] [Amgoud et Prade, 2012] [Bench-Capon, 2012] [Gomez et al, 2013]. Quelques travaux empiriques font appel à l'étude de l'argumentation dans les documents [Mann et Thompson, 1988] [Teufel et al, 1999] [McBurney et Parsons, 2000] [Adelman et al, 2007] [De Jong , 2008] [Hartley et Betts, 2008]. Ces travaux se penchent davantage sur la recherche d'une structure globale du document. Des études fines et intéressantes sur le discours vont au-delà de l'analyse linguistique en élaborant des approches automatiques d'annotations du discours jusqu'à des arguments comme la notion de résultat ou de contraste [Baldridge et al, 2007][Webber et Prasad, 2009][Braud et Denis, 2013]. Notre approche part d'un corpus du discours scientifique d'un domaine pour obtenir un modèle de langage (que l'on espère assez général mais peut être propre à ce domaine) a contrario de ces approches qui s'inspirent d'un modèle existant à partir d'un corpus étalon comme le RST DT (voir 3.2.). Un résumé de projet est un document assez contraint, les auteurs devant préciser un contexte (scientifique et/ou sociétal), des objectifs, des moyens d'y parvenir, et des résultats et impacts escomptés, le tout en soulignant l'originalité. Notre étude, présenté dans cette article, rebondit sur ces éléments, plus proche de la structure du langage, que de la structure d'un document, pour caractériser des motifs typiques de ce genre de document.

[Lu et al, 2011] établissent une étude très proche de la nôtre par comparaison de l'emploi récurrent d'explications dans les réactions d'un groupe d'étudiants dans un forum collaboratif. Mais à l'inverse de notre approche ils utilisent un précodage des informations du corpus, alors que nous cherchons à découvrir des motifs d'argumentation. [Fréard et al, 2010] ont étudié les conflits d'un débat épistémique sur Wikipedia dans le domaine de l'astronomie pour comprendre l'influence des processus argumentatifs dialogiques dans les échanges socio-relationnels. Ils reconnaissent que la connaissance argumentative est tacite et difficile à modéliser par des traitements automatiques ; ici encore, la plupart des traces d'argumentation sont codées à la main. Notre travail s'inscrit à la frontière de ces travaux à l'échelle d'un corpus et de ceux de [Budzynska et Reed, 2011][Walton, 2012] qui considèrent l'induction locale de motifs d'argumentation sans aller, comme ici, jusqu'au balayage des motifs à l'échelle d'un corpus.

Après avoir illustré la genèse et les objectifs de notre projet, nous ferons un rappel théorique sur l'argumentation. Nous passerons ensuite à la présentation des méthodes et des outils mobilisés dans cette étude, avant d'illustrer quelques résultats et les perspectives que ce travail ouvre.

2 Position et corpus

2.1 Position

Sans ignorer l'existence d'un large ensemble de travaux scientifiques portant sur la mesure de la production scientifique, ni ceux plus actuels attachés à la cartographie des sciences, nous positionnons notre travail dans le domaine de l'analyse et de la visualisation de données textuelles produites par les scientifiques et réunies en corpus calibré. Ce domaine est reconnu pour sa contribution à la réflexion sur la composition et l'évaluation des politiques scientifiques [Callon et Law, 1986]. Aujourd'hui, l'évolution de l'analyse des dynamiques scientifiques est fortement portée par la question de la caractérisation des collaborations et des dynamiques cognitives de la production de savoir [Powell et al., 2005] et par l'émergence de domaines de recherche multi ou transdisciplinaires [Lucio-Arias et Leydesdorff, 2007] ou par la reconstruction de l'évolution de champs scientifiques. Tracer et cartographier la structure et l'organisation des connaissances représente un enjeu important pour plusieurs disciplines concernées par les questions d'extraction d'information. Il en va de même pour les travaux en sociologie des sciences, pour lesquelles la compréhension des dynamiques sociales et cognitives des activités de recherche est un point de passage important [Cambrosio et al. 2004; Cambrosio et al. 2006; Bourret et al. 2006; Bonaccorsi 2008]. Cela convoque évidemment des débats importants. [Zitt and Bassecoulard 2008] ont ainsi indiqué trois enjeux importants pour les études scientométriques : 1) la qualité de la description des dynamiques de savoirs est dépendante de plusieurs sources de données; il est aussi important de 2) caractériser ces dynamiques que 3) d'évaluer les positions des producteurs en réduisant les problèmes posés par la variété.

Notre compréhension de ces enjeux - qui accompagnent certaines critiques que nous partageons sur la puissance et l'intérêt des indicateurs de production [Barré 2001; Freeman et Soete 2009], nous conduit à orienter notre travail vers l'étude des collaborations scientifiques qui s'établissent dans les projets de recherche financés sur appel d'offre [Heimeriks et al., 2003]. C'est un objectif important puisque la forme « projet » est devenue un mode d'organisation de collectifs de recherche du fait de la généralisation du financement de la recherche par appel d'offre, instrument d'incitation majeur pour la politique des sciences et des techniques. En prenant le projet en lieu et place de la publication scientifique comme objet, nous pensons avoir accès à une dimension sous-étudiée des collaborations scientifiques et techniques, appréhendées du coup aussi comme des formes d'engagement de l'activité de recherche dans le futur, ce qui est aussi une propriété très intéressante. Notre travail rejoint une voie de recherche collective présente en France sur l'auctorialité scientifique qui, après des travaux importants de sociologie pragmatique de la signature scientifique pour accéder aux différentes conceptions de l'auteur qu'elles contiennent [Pontille, 2004], propose de constituer une étude sociale des modes d'expression variée de l'auteur scientifique en lien avec une approche des formes linguistiques ou sémiotiques des discours qui constituent l'auteur [Grossmann, 2010] [Tutin et Grossmann, 2014].

2.2 L'importance de l'étude des collaborations dans les projets de recherche

L'article de [Katz and Martin, 1997] est un des articles fondateurs de cette perspective, auquel l'agenda de recherche proposé par [Beaver, 2001] a fait écho pour justifier l'importance des collaborations de projets comme objet d'étude. [Katz and Martin, 1997] ont explicitement fondé leurs analyses sur des travaux empiriques, empruntant leurs méthodes à la scientométrie et focalisant leur attention sur les dimensions de la productivité des chercheurs lues à différents niveaux (voir ibid., pp 6-14). Les travaux empiriques de [Beaver, 2001: 372-375] ont davantage porté sur les motivations à la collaboration, en établissant des motifs variés, non centrés sur celui de la publication. Malgré les résultats des travaux de [Bozeman and Rogers 2002] qui les conduisent à affirmer un peu rapidement que « *beaucoup de scientifiques ne conçoivent pas leur travail en termes de source de financement ou d'évaluation de projet, mais qu'ils considèrent plutôt les projets comme de purs artifices bureaucratiques*» (ibid., p.771), nous pensons que les textes qui fondent ces projets contiennent des informations pertinentes sur le sens donnés par les chercheurs à leurs pratiques de recherche [Prieto, 1975] et à la structure relationnelle de leurs activités professionnelles [Lazega, 2011]. On ne peut donc considérer que les textes de projet soient uniquement d'arguments de façade. Il nous semble a minima que l'usage stratégique des collaborations exposées au sein des projets représente une véritable question empirique et que les contenus textuels de ces projets doivent faire l'objet d'une étude précise et équipée. Certes, une telle approche doit s'accompagner d'un travail d'enquêtes sociologiques spécifiques à l'étude de ces structures relationnelles. Un tel objectif est renforcé par le fait qu'un travail aux frontières des programmes, des agences et des instituts de recherche est aujourd'hui reconnu comme une des arènes où s'élaborent les orientations scientifiques [Guston, 2001]. Enfin, il faut souligner que l'approche projets s'avère parfois plus efficace pour cibler un domaine scientifique protéiforme (tel que celui de la biodiversité) : là où une liste de mots clés est très difficile à établir pour ce ciblage, la sélection des appels d'offre et projets pertinents peut s'avérer très efficace [Chaveriat et al, 2011].

2.3 Une attention portée au domaine de recherche sur la biodiversité

Notre entreprise scientifique peut concerner un grand nombre de domaines de recherche. Nous avons cependant jugé utile de nous consacrer à l'étude d'un domaine de recherche émergent en faisant l'hypothèse que les stratégies et objectifs scientifiques seraient plus nettement marqués par l'affirmation de contenus jugés « innovants » ou porteurs d'agencements nouveaux par les chercheurs qui déposent ces projets. Nous suivons en cela certaines des recommandations établies par [Bonaccorsi 2008] concernant les domaines scientifiques d'exploration qui à la fois contiennent l'affirmation de nouveaux objets de recherche sans un marquage nécessairement disciplinaire, et des processus d'institutionnalisation fondés sur de grands programmes sur appel d'offre. Dans cette perspective, il devenait pertinent pour nous de mobiliser les travaux réalisés dans le projet PanBioptique financé par l'ANR sur la question de l'institutionnalisation de la biodiversité, et de collaborer avec le réseau européen Biodiversa (http://www.biodiversa.org) qui a récemment développé une base de données réunissant des informations sur plus de 6000 projets de recherche sur la biodiversité.

Nous avons ainsi défini un objectif commun de fouille de texte dans une base de données de notices décrivant les projets de recherche financés, et porté notre attention sur ces textes courts que sont les résumés. L'objectif était donc de tirer avantage de l'important travail réalisé par les membres de Biodiversa pour constituer une base de données riche et bien renseignée, et de développer – de façon collaborative – une perspective d'extraction d'information au niveau de la sémantique en usage dans ce genre de texte. En effet les textes de résumé de projets sont assez particuliers car ils tentent un agencement délicat entre différents registres de justification et de conviction pour rendre un espace de collaboration putatif attrayant et clair : ils essaient d'affirmer dans l'espace du programme de l'appel d'offre la promesse dont sont porteurs les collaborateurs.

2.4 Les objectifs de l'analyse des résumés de projet de la base de données Biodiversa

Notre projet s'enracine sur un terrain à forte vocation interdisciplinaire : la sociologie des sciences d'un côté, la linguistique – représentée par différents domaines : l'analyse du discours, la pragmatique, la terminologie – et l'informatique – notamment, en ce qui concerne la fouille de textes, le traitement automatique des langues et l'extraction d'information – d'un autre. En raison de cette interdisciplinarité, nous avons fixé divers objectifs à atteindre, qui se situent chacun dans le cadre d'une ou plusieurs disciplines.

Les corpus – que nous décrirons en détail plus loin – sont composés de résumés scientifiques en langue anglaise et décrivent des projets ayant trait à des secteurs variés de la biodiversité en vue d'obtenir des financements. Il s'agit donc de textes argumentatifs qui sont intéressants à analyser au prisme des disciplines citées plus haut. La détection des motifs d'argumentation mobilisés dans les résumés constitue l'objectif principal du projet : comprendre comment les auteurs des résumés bâtissent leur discours pour convaincre les destinataires a été la première direction dans laquelle nos recherches ont été orientées. On comprendra tout de suite l'intérêt d'une telle étude : une fois ces motifs d'argumentation identifiés, il sera possible de les tester sur d'autres corpus spécialisés pour vérifier s'ils correspondent à des motifs d'argumentation généraux et non spécifiques aux corpus étudiés dans notre projet. Ces motifs d'argumentation sont décrits sous forme de motifs. La création d'une base de motifs propres à l'argumentation dans les textes scientifiques est un objectif connexe à notre objectif primaire.

Outre les objectifs que nous venons de citer, nous souhaitions déployer une attention particulière à des notions caractéristiques des domaines de recherche sur la biodiversité de façon à aborder des motifs généraux d'argumentation en fonction de notions prépondérantes dans les résumés. Un autre objectif est l'élaboration de motifs pour détecter des entités nommées : noms d'espèces, taxons, noms d'organisations et de technologies. Cette étude mobilise pour ce faire les cadres des théories de l'argumentation qui s'avèrent particulièrement adaptés à expliquer notre approche.

3 Etat de l'art sur l'étude des arguments

3.1 Théories générales de l'argumentation

L'étude de la littérature de l'argumentation a été une étape capitale dans notre projet. Le but a été d'identifier, parmi les nombreuses théories existantes, celles qui convenaient à l'analyse des résumés de projets, tenant compte de ce genre "littéraire" particulier qui vise une présentation du contexte, des buts et des moyens d'une activité scientifique projetée décrite dans un document sous-jacent. Voilà la définition de l'argument utile à notre approche.

Définition. *Argument*
Raison ou ensemble de raisons données en support d'une idée, d'une action ou d'une théorie.

Dans un domaine comme la biodiversité, domaine de recherche scientifique et technologique, l'argument peut recouvrir une vaste variété de formes : un nom d'espèce, un objectif, une technologie utilisée, une mesure chimique à appréhender. C'est vaste et notre hypothèse se base sur le fait qu'il est possible de s'appuyer sur des marqueurs (prépositions, verbes, adverbes) et combinaisons de marqueurs pour détecter la présence d'un argument.

La définition donnée par Charaudeau et Maingueneau (2002) dans leur Dictionnaire d'analyse du discours reprend les apports de différentes théories sur cet objet d'étude. Tout d'abord, les auteurs affirment que « le discours argumentatif a été caractérisé de façon intradiscursive par ses différentes formes structurelles et, de façon extradiscursive, par l'effet perlocutoire qui lui serait attaché, la persuasion. » [2002 : 66]. L'effet perlocutoire dont il est question dans cette citation est bien l'étude des techniques discursives employées pour provoquer ou augmenter l'adhésion du lecteur/récepteur aux thèses de l'auteur/émetteur, suivant les travaux de C. Perelman et L. Olbrechts-Tyteca (1970).

Pour ce qui concerne le terrain du domaine de l'argumentation, si dans l'Antiquité il était restreint aux genres traditionnels de la rhétorique, de nos jours il coïncide plus généralement avec le genre du débat sous toutes ses formes.

Deux distinctions principales de l'argumentation sont opérées par Charaudeau et Maingueneau : d'un côté, ils définissent l'argumentation comme la présentation d'un point de vue, qui peut prendre plusieurs formes (argumentation en plusieurs énoncés, argumentation en un seul énoncé, argumentation en un seul mot) ; d'un autre, ils la présentent comme « mode spécifique d'une constellation d'énoncés », sachant que les deux définitions ne s'excluent pas mutuellement [2002 : 67].

Bien que les théories de l'argumentation soient à ce jour plutôt nombreuses, elles ont souvent été développées en prenant comme objet d'étude des discours oraux ou écrits pour être prononcés à l'oral, notamment dans le domaine juridique ou politique. Nous n'avons vraiment pas trouvé de modèle théorique parfaitement compatible avec notre objet d'étude. Toutefois, nous avons pu emprunter quelques notions à certains des modèles théoriques passés en revue dans l'étude de la littérature.

Ainsi, pour n'en citer que quelques-uns, nous avons focalisé notre attention sur les notions suivantes. Tout d'abord, l'étude des connecteurs dans l'ADL (Argumentation Dans la Langue) de [Ducrot et Anscombre 1980 et 1983]. Les connecteurs sont des mots stratégiques

dans la construction d'un discours de type argumentatif : ils sont indispensables pour introduire et agencer les arguments, et pour discuter des causes et des conséquences de certains sujets. Nous avons remarqué que les résumés de projets constituant nos corpus font largement usage de ces unités linguistiques.
Ensuite, le modèle de la logique substantielle de Toulmin nous a semblé intéressant pour ce qui concerne la structure des textes analysés. [Toulmin 1958] identifie cinq composantes fonctionnelles dans tout passage argumentatif : des données [Data], une thèse ou une conclusion [Claim], une loi de passage ou garant [Warrant], qui sert à assurer le lien entre les deux premiers éléments et qui, en même temps, provoque une régression potentielle à l'infini, représentée par un support [Backing]. Le dernier élément du passage argumentatif de ce modèle est le modalisateur [Qualifier], qui est souvent un adverbe et définit une restriction [Rebuttal]. La cellule monologique argumentative est définie par sa structure : un locuteur avance une thèse ou conclusion [Claim] en l'appuyant sur une donnée [Data] et sur des règles [Backing, Warrant]. Cette thèse peut être réfutée sous certaines conditions [Modal, Rebuttal]. Le fait que nos abstracts débutent presque toujours par une ou deux phrases qui servent de [Data], nous a conduit à prendre en considération ce modèle théorique pour notre étude.
[Eemeren, van, et Grootendorst 1992] ont développé une théorie générale de l'argumentation qu'ils ont appliqué à l'argumentation du discours politique, du discours inter-personnel ou la communication sur la santé. Leur théorie met l'accent sur des aspects dialectiques et pragmatiques et voit l'argumentation comme une mise en oeuvre complexe d'actes du discours résultant d'activités en langage naturel d'une part, et pourvu d'objectifs de communication spécifiques d'autre part**.** Un des fondements a reposé sur une articulation de l'argumentation comme moyen de résoudre des différences d'opinion. L'argumentation commence avec quatre principes 1) L'externalisation: l'argumentation a besoin d'une attitude et d'une opposition à cette attitude. Ainsi la recherche d'arguments se focalise davantage sur les engagements que sur l'état psychologique des acteurs. 2) La socialisation: les arguments sont vus arguments comme l'expression de processus entre acteurs. Il est important de valider la position d'un acteur par des arguments d'une certaine manière. Deux personnes essayent d'obtenir un accord dans l'argumentation ; ainsi l'argumentation est une partie d'un contexte social qui dépasse le contexte individuel. 3) La fonctionnalité: l'argumentation possède une fonction générale de gestion de résolution d'un désaccord. L'étude d'une argumentation doit se concentrer sur la fonction de l'argumentation dans la gestion verbale d'un désaccord et finalement 4) La dialectification: l'argumentation est appropriée seulement quand on est capable d'utiliser des arguments qui sont aptes dans l'aide à argumenter contre un autre acteur. Dans cette théorie l'analyse d'une argumentation se réalise selon des régularités, des règles et des échanges pragmatiques sur le gain ou non dans une situation.

3.2 Théories de l'argumentation pour le discours technique

[Ivin, 2000] a exploité la théorie de [Eemeren, van, et Grootendorst 1992] dans le débat scientifique et notamment à travers sa production. Il a dégagé des types d'arguments utilisés distribués dans un corpus selon la classification suivante: l'argumentation empirique, l'argumentation théorique, l'argumentation contextuelle, l'argumentation épistémologique.
La RST (Rhetorical Structure Theory) de [Mann et Thompson 1988] nous a semblé un modèle théorique digne d'intérêt car il s'appuie sur l'analyse de textes écrits. La théorie RST est un cadre d'analyse du discours bien étudié. Dans la RST, un morceau de texte est découpé en

une séquence de fragments non recouvrants appelés unités de discours élémentaire ou EDU (elementary discourse unit). Des EDU voisins sont reliés entre elles par une relation typée. La plupart des relations de la RST sont qualifiées de hypotactiques, c'est-à-dire qu'une des deux EDU participant à la relation est démarquée comme noyau et l'autre comme satellite. Le noyau est plus important du point de vue du rédacteur, tandis que le satellite apporte plus d'information pour aider la compréhension du noyau. Certaines relations sont paratactiques, dans celles-ci les deux EDU sont noyaux. Dans la RST un arbre du discours est consitué en considérant chaque EDU comme des noeuds feuilles. Les noeuds dans l'arbre du discours sont associées les unes aux autres par des relations qui relient les EDU. Les relations du discours de la RST capturent les relations sémantiques entre des EDU. Par exemple de telles relations offrent des indices de relations temporelles entre des événements entre deux EDU. Sur le site de l'université de Pennsylvanie (Linguistic Data Consortium), on peut acheter un corpus annoté selon la théorie de [Mann et Thompson 1988] pour établir des évaluations (RST Discourse Treebank). Il réunit 385 documents du Wall Street Journal annotés par des étiquettes rhétoriques. En moyenne un document est couvert par 57 annotations. Si toutes les catégories ne concernent pas directement notre objectif d'identifier des éléments de persuasion dans un projet, on trouve des catégories intéressantes comme « verbes cognitifs ».
[Cabrio et Villata, 2012] n'ont ni développé de théorie spécifique à l'argumentation, ni utilisé le RST Discourse Treebank, mais ont adapté un cadre de repérage d'argument pour l'étude de controverses sur internet. Leurs travaux donnent une primauté à la notion d'argument et à son extraction dans les documents (réseaux sociaux, forums…). Selon leur définition un argument est une opinion formulée par un participant à un débat (ou controverse) qui vient consolider une opinion existante ou la contredire. Cette opinion se formule par une phrase et les phrases en opposition ou non se suivent dans leur contexte avec une similarité lexicale (TE ou *text entailment*). Notre idée d'argument dépasse la notion d'opinion dans le sens où plusieurs catégories cognitives ou du monde réel peuvent être mises en jeu sur un contexte scientifique. Ainsi un nom d'espèce (étude des espèces en voie d'extinction), une localisation (le littoral) ou une hypothèse (impact du changement climatique sur une population d'organismes vivants) peuvent devenir des arguments à part entière.

3.3 Extraction lexicale dans les textes

La création d'un dictionnaire terminologique à partir d'un corpus est une problématique ancienne et bien étudiée en traitement automatique des langues [Smadja et McKeown, 1993][Enguehard et al, 2002]. L'extraction du lexique de mots simples (i.e. monotermes ou 1-grammes) est une opération qui se fait grâce à une segmentation des phrases à l'aide d'une liste de délimiteurs qui est souvent suivis d'une opération de tri par fréquence pour sélectionner les éléments les plus intéressant que l'on filtre par une liste de mots outils. Cette opération ne pose pas de question technique ni scientifique. L'extraction de mots composés peut s'avérer plus délicate s'il s'agit de détecter des entités multi-mots (i.e. multitermes ou n-grammes) contigus, non contigus, intégrant des verbes ou des prépositions et au moyen de les filtrer pour en retenir les plus pertinents par rapport à un objectif particulier. Dans notre cas si on imagine l'extraction de marqueurs comme « a pour effet … sur … » l'extraction de multitermes non contigus peut s'avérer intéressant. L'extraction de multitermes contigus incluant des verbes comme « notre hypothèse montre que » est typiquement un segment répété dont l'extraction est indépendante du typage syntaxique (déterminant, verbe, noms, adjectif..) des mots simples qui le constitue. L'extraction de segments répétés [Lebart et al,

1998] ou de multitermes non contigus [Doucet et Ahonen-Myka, 2006] génère un espace de candidats très importants de plusieurs dizaine de milliers de candidats qu'il est couteux de trier. C'est le souci de ces techniques qui nécessite un effort d'ingénierie de nettoyage (en anglais *database curation*) très important.

4 Méthodologie

4.1 Approche globale

Notre démarche est prospective dans le sens où nous cherchons une théorie du langage (i.e. ensemble de motifs voir définition ci-dessous) permettant de mieux comprendre la structure des documents. Nous avons n'avons adopté aucun parti pris sur les théories d'argumentation et les arguments à rechercher selon ces théories pour ne pas entacher l'identification empirique de types d'arguments. L'identification s'est faite selon la lecture visuelle et linguistique de plusieurs centaines de projets et selon notre pratique d'auteur scientifique. Cette démarche « corpus-driven » procède donc à l'élaboration empirique de règles selon une logique propositionnelle (combinaison de conjonctions d'éléments lexicaux). Une autre démarche, « corpus-based », aurait été comme (Fang et Hirst, 2012) d'adopter une méthode automatique d'annotation par apprentissage selon un corpus étiqueté d'après une théorie d'argumentation comme la RST. Notre démarche nous permet de détecter (si elles existent) d'éventuelles règles non présentes dans une telle théorie.

Définition. *Motif d'argumentation*
On appelle motif d'argumentation une séquence d'un ou plusieurs mots servant de marqueur pour annoncer un argument. On peut aussi qualifier le motif de grammaire locale. Un tel motif doit pouvoir se construire par un automate à états finis.

Cette théorie du langage se définit par des éléments qui illustrent la présence d'un argument scientifique. Il est donc difficile de valider l'approche sachant que la théorie peut être incomplète et nous n'avons pas la connaissance des limites de l'espace de recherche pour établir cette théorie du langage. On peut quand même en jouant sur l'échantillonnage avoir une estimation de la qualité des motifs.

La figure 1 décrit les différentes étapes de notre démarche globale. La première étape a permis de réfléchir la construction du corpus. Elle a duré plusieurs mois car la base actuelle a bénéficié d'une mise à jour importante à compter du mois de juin 2013 date à laquelle nous avons travaillé sur un corpus de taille optimale couvrant la majorité des résumés de projet acceptés aux appels d'offre sur la biodiversité à l'échelle européenne. En 2012 nous n'avons bénéficié que de la moitié de ce corpus.

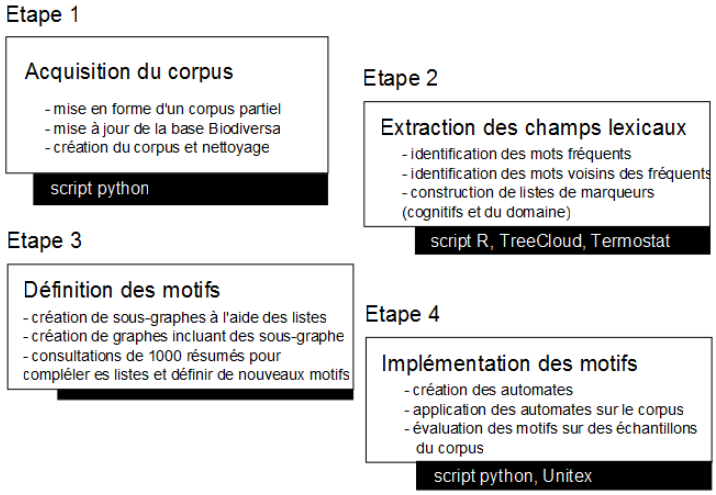


FIG. 1 – *Approche globale de création et d'application des motifs d'argumentation*.

L'étape 2 a permis d'identifier des éléments lexicaux clés intrinsèques au corpus. L'hypothèse fondamentale repose sur le fait que les motifs apparaissent fréquemment dans les documents du corpus, donc ils sont formés par des éléments lexicaux fréquents et repérables dans des listes de mots fréquents et d'association entre ces mots fréquents et d'autres mots plus ou moins fréquents. Les outils utilisés Treecloud, Termostat et la librairie TM de la plateforme R ont été utilisés à cette fin et répondent parfaitement à l'identification de telles formes lexicales. L'étape 3 est manuelle et correspond à la définition des motifs par consultation des résumés grâce aux lexèmes identifiés à l'étape précédente. Comme l'étape 1, cette étape est très couteuse en temps passé, de l'ordre de plusieurs mois. Finalement l'étape 4 repose sur l'implémentation des motifs sous forme d'automates à états finis dans l'outil Unitex et l'exploitation des exports au format html pour réaliser des évaluations.
Cette démarche a été spécifiée avec le corpus Biodiversa de projets sur la biodiversité. Cependant notre démarche présentée figure 1 a l'ambition de se généraliser à d'autres corpus calibrés de projets sur d'autres domaines comme les nanotechnologies, les énergies durables, la chimie verte, les technologies de l'information, les biotechnologies, l'économie participative.

4.2 Présentation des corpus

La base de données de projets européens sur la biodiversité de laquelle nous avons extrait notre corpus a été développée de 2005 à 2013 par le consortium BiodivERsA. Cette base a été constituée en important et sélectionnant les données provenant de plusieurs agences de financement, ainsi que celles dérivées d'appels à projets européens. Une attention particulière a été portée sur les thématiques des projets, afin de n'inclure que ceux en rapport, au moins partiellement mais significativement, avec la biodiversité. La base de données utilisée représente un total de 603 appels à projets, dont 100 spécifiques à la biodiversité, parus dans 17 pays ou au niveau européen.

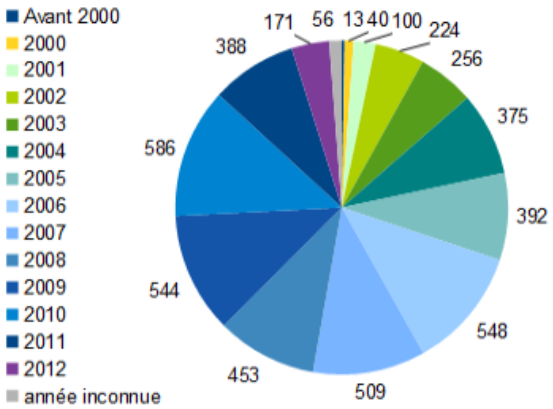


FIG. 2 – *Répartition des projets de recherche extraits de la base Biodiversa, en fonction de l'année de début de projet.*

Nous avons exploité un ensemble de 4655 projets rédigés en anglais dont le résumé est disponible. La répartition chronologique de ces projets en fonction de l'année de début de projet est donnée en figure 2. Ce corpus représente un total d'environ 1,4 millions de mots.

4.3 Identification des éléments de base des motifs ou sélection d'attributs

Dans notre hypothèse d'analyse, on suppose que des motifs d'argumentation stables jouent un rôle dans la présentation d'un projet. La stabilité signifie la répétition modulée par une variabilité terminologique. Une approche robuste et rapide pour capturer la répétition consiste à extraire les mots simples, racinisés ou pas, et fréquents. En première approximation, on peut considérer une fréquence seuil par rapport au nombre de projets. Une partie des mots fréquents (substantifs, verbes ou adjectifs) ont une implication dans la définition manuelle des motifs associés à l'argumentation.

4.3.1 Distribution globale des mots simples

L'étude distributionnelle des mots fréquents permet de capturer un signal fort sur les occurrences les plus significatives sous l'hypothèse de leur répétition. Le travail précurseur de Georges Zipf a montré une distribution typique x.y = Constante où x est le rang en termes du nombre d'occurrences d'une forme simple dans un texte (ou un ensemble de textes) d'une langue donnée et y le nombre d'occurrences de cette forme simple [Zipf, 1935]. La fréquence est définie par le nombre d'occurrences d'une forme dans le corpus.

Comme notre étude concerne la définition de motifs d'argumentation qui couvrent un pourcentage important de documents, l'extraction de mots simples fréquents peut s'avérer être une bonne source de marqueurs. On peut la qualifier d'étape de sélection de variables ou sélection de caractéristiques selon la terminologie du domaine de l'extraction des connaissances. [Hernandez et Grau, 2003] ont utilisé une méthode de filtrage intéressante à partir des mots fréquents de la langue générale pour identifier les mots saillants d'un domaine. Dans notre cas, il s'agit d'identifier des patrons qui peuvent souvent être construit avec des

mots de langue générale comme « our hypothesis shows that » dont les quatre mots simples apparaissent dans la liste des 5000 mots les plus fréquents du British National Corpus.

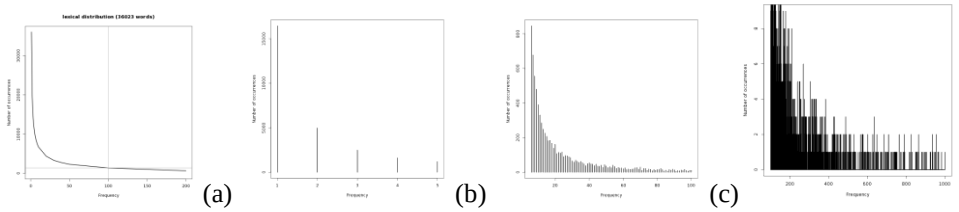
Pour la sélection de variables nous avons utilisé le package tm qui offre des fonctionnalités de gestion des documents textuels, et d'abstraction du processus de manipulation de document et facilite l'usage de formats textuels hétérogènes dans R [Feinerer et al, 2008]. tm fournit un accès facile aux mécanismes de prétraitement et de manipulation tels que la suppression de la ponctuation et des blancs, la racinisation, ou la suppression de mots vides. En plus une architecture de filtrage rend possible la sélection de certains documents selon un certain critère et procède à une recherche plein texte. Le package propose l'export de matrices terme-document à partir d'une collection de documents.

Texte brut	Texte racinisé
The project envisages to continue and extend the studies of evolution and systematics in the grass genus Bromus and the legume genus Vicia supported by our ending Estonian Science Foundation Grant No.4082 for 2000-2003. The project combines traditional morphology-based botanical systematics, biochemical isozyme analyses, genetics, and chromosome cytology for solving problems of phylogenetic systematics, phylogeography and evolution in botany. The main objectives of the project are: 1. To determine genetic divergence and relationships within and among species Bromus hordeaceus, B. secalinus, B. racemosus and B. commutatus of type section of genus Bromus by cladistic and phenetic analysis of isozymes with checking the correspondence of the results with the traditional morphological species delimitation	*project envisag continu extend studi evolut systemat grass genus bromus legum genus vicia support estonian scienc foundat grant project combin tradit morphologybas botan systemat biochem isozym analys genet chromosom cytolog solv phylogenet systemat phylogeographi evolut botani main object project determin genet diverg relationship speci bromus hordeaceus secalinus racemosus commutatus type section genus bromus cladist phenet analysi isozym check correspond result tradit morphology speci delimit*

TAB. 1 – *Exemple de texte brut correspondant à une partie d'un résumé d'un projet (à gauche) et texte racinisé obtenu après traitement (à droite).*

Le corpus contient 36023 mots simples dont 19588 de fréquence supérieure à 2 (voir figure 3). 4966 formes apparaissent exactement 2 fois, 2509 formes apparaissent exactement 3 fois. Si on opère une racinisation (i.e. stemming ou troncature de mots conservant le radical), comme le montre la transformation du tableau 1, on obtient 25875 formes dont 13519 ont une fréquence supérieure à 2. 3568 formes apparaissent exactement 2 fois, 1734 formes apparaissent exactement 3 fois. Le taux de réduction de la variabilité des mots est de 29% pour l'intervalle de fréquences [2-3].

Que ce soit avec ou sans racinisation, les mots de fréquence égale à 1 représentent entre 45.6 et 47.78 % des formes, les mots de fréquence égale à 2 ou 3 représentent entre 38.1 et 39.2 % des formes de fréquence supérieure à 1, ou entre 20.4 et 20.5 % des formes en général (voir figure 3).



(a) (b) (c) (d)

FIG. 3 – *Fréquences cumulées des mots simples bruts pour toute fréquence (a). Distribution des occurrences des mots utilisés dans le corpus en fonction de l'intervalle de fréquences : [1-5] (b), [6-100] (c), [101-10000] (d).*

4.3.2 Distribution globale des mots simples

Pour extraire des descripteurs significatifs, un procédé rudimentaire et classique consiste à trier par ordre de fréquence et retenir les plus fréquents. Il est intéressant, dans le cadre de cette étude sur un espace de document-projets, de se référer à la présence des descripteurs dans l'espace des documents. On opère une sélection par seuil de couverture de l'espace de documents. Le seuil de couverture est une mesure simple qui représente une fréquence brute d'usage. D'autres mesures pourraient être utilisées comme la sélection par clustering des mots associés, ou l'utilisation d'une mesure pondérée comme le tf.idf (*term frequency-inverse document frequency*). Dans un cas on tient compte des contextes communs ce qui force à retenir des mots partageant des contextes similaires ; dans l'autre cas on retient des mots fréquents mais typiques parce que fortement distribués dans un faible nombre de contextes (occurrences élevées au sein d'un même document ou un sous-corpus). Dans notre cas ces contraintes sont trop fortes et on risque de perdre beaucoup de termes. D'autant plus que le corpus n'est pas considérable si on le compare au nombre de documents (plusieurs millions) que l'on trouve dans des corpus-étalons.

Définition : Seuil d'usage.
On appelle Sf le seuil d'occurrences d'un mot tel que :
Sf = Sd . D
où D est le nombre de documents du corpus et Sd un pourcentage de documents. On fixe un seuil bas pour les mots fréquents (Sd_b) et un seuil haut pour les mots ultra-fréquents (Sd_h).

On fixe, intuitivement, Sd_b à 5% comme seuil bas de quantité de documents pour qu'un descripteur soit représentatif en termes d'occurrence. Avec un tel seuil le nombre de mots avec Sf =232 pour notre corpus. On remarque que le nombre de mots fréquents est faible. L'espace de descripteur est de 569 mots fréquents bruts soit environ 1,5% des mots bruts, et 617 mots fréquents racinisés. Il peut être assez rapidement consultable visuellement, ce qui est d'autant plus pratique quand on veut se rendre compte d'un contenu et procéder à une méta-analyse.

Le tableau 2 permet de consulter la liste des mots ultra-fréquents, c'est-à-dire apparaissant dans le corpus au seuil haut Sd_h de 50% soit un seuil d'occurrences Sf =2727.

Mot racinisé	Fréquence	Mot racinisé	Fréquence	Mot brut	Fréquence

studi	5529	climat	2780	different	3105
chang	5212	can	2757	study	3049
project	4893	genet	2752	can	2757
popul	4225	import	2743	biodiversity	2655
environ	4065	research	2692	also	2574
differ	3975	biodiv	2665	data	2447
ecosystem	3335	also	2574	research	2444
develop	3305	data	2448	change	2408
model	3193	ecolog	2415	genetic	2390
plant	3184			changes	2364
effect	3132			populations	2340

TAB. 2 – *Liste des mots ultra-fréquents au seuil Sf=50% du nombre de documents du corpus.*

Une liste de 488 mots vides (comme *yourselves, already*…) a servi de filtre ; d'autre part les mots de taille inférieure à 3 caractères ont été négligés. Cela explique l'invisibilité des verbes auxiliaires *have* et *be*. Il y en a assez et les balayer permet de se rendre compte du contenu des documents-projet en quelques secondes. La granularité descriptive est évidemment, à ce stade, grossière.

4.3.3 Visualisation globale des attributs pertinents

Pour pouvoir construire des règles pertinentes il faut inclure des éléments lexicaux pouvant être des noms mais aussi d'autre type. Dans cette analyse nous essayons de faire une projection des associations entre mots simples moyennement fréquents indépendamment de leurs catégories. Il ne s'agit pas de visualiser des collocations, c'est-à-dire des mots fortement associés entre eux dans une phrase, qui à l'extrême formeraient des expressions figées ; mais il s'agit de visualiser des mots apparaissant dans un même contexte. En construisant une matrice d'association avec pour fenêtre de cooccurrence la taille d'un résumé, le taux de « sparsité » (état creux) atteint plus de 90%. Si nous restreignons la fenêtre de cooccurrence à quelques mots nous perdons un signal associatif fort sur l'espace global des termes. Cela devient peu pertinent pour une visualisation globale mais surement plus pour une visualisation locale comme avec l'outil TreeCloud présenté plus loin. Notre objectif consiste donc à visualiser les liens de voisinage de l'espace globale pour pouvoir projeter une catégorie lexicale en particulier et apprécier l'intérêt de la prendre en compte. Typiquement nous allons le faire pour une liste de verbes. Car nous supposons que certains motifs peuvent inclure des verbes comme « our results show that » plus pertinent que « results » tout seul.
Pour la classification globale et la visualisation nous avons utilisé les packages de base et le package Igraph qui est un package d'analyse des réseaux et de visualisation [Csardi et Nepusz, 2013]. Treize algorithmes d'affichage sont disponibles. Nous avons utilisés deux algorithmes « dirigé par la force ». Ce type de visualisation est compatible avec le traitement d'un grand graphe dans lequel les nœuds sont vus comme un système physiques sur lequel on applique une force (attraction ou répulsion) repoussant ou attirant les nœuds deux à deux. Nous avons utilisé le placement de Fruchterman-Reingold dont les forces influençant une paire de nœuds sont la loi de Hooke pour l'attrction et la loi de Coulomb pour la répulsion [Fruchterman et Reingold, 1991]. Nous avons aussi utilisé l'algorithme DrL (Distributed

Recursive Layout) qui est “dirigé par la force » et utilise un algorithme basé sur la routine VxOrd en proposant une version multi-niveau récursive pour obtenir un meilleur placement sur de grands graphes, avec aussi la capacité d’ajouter un nouveau nœud à un graphe déjà dessiné [Martin et al, 2011].

Comme précédemment, cette partie est une étude des distributions à grande échelle. Un motif (telle que cela sera abordé plus loin) peut être vu comme un graphe d’associations locales, donc entre voisins. Si on essaye de visualiser le graphe global des associations locales, les éléments clés qui apportent les éléments des motifs doivent aussi figurer sur ce graphe global. On peut donc se servir d’un graphe global pour tester l’importance d’un ensemble d’éléments dans la globalité. La visualisation, proposée ici, consiste à s’attacher sans connaissance structurale préalable à discerner l’importance, ou non, de l’utilisation des verbes dans des structures associatives globales. Notre démarche ici s’appuie sur une technique de classification des proches voisins associée à une technique de visualisation ‘dirigée par la force’. Les deux techniques sont globales et doivent s’attacher à considérer toutes les formes fréquentes. Les verbes sélectionnés pour contribuer aux motifs sont projetés pour observer leur contribution dans les composantes connexes.

L’algorithme des proches voisins [Cover et Hart, 1967] a donné de bons résultats pour la classification ; un nouvel objet est classifié par la majorité des votes de ses proches voisins grâce auxquels la classe majoritaire est attribuée. C’est un algorithme d’apprentissage supervisé pour affecter une classe à un individu inconnu en fonction de l’appartenance de classe de ses voisins, mais on peut l’utiliser comme algorithme d’apprentissage non-supervisé pour afficher ou extraire des voisinages notamment. Il converge rapidement et tient compte des contextes proches (collocations).

Soit la matrice de données $M = (u_{ij})_{1 \le i \le n,\ 1 \le j \le m}$ où i représente la i-ième ou mot i ; j représente la j-ième colonne ou document j ; n est le nombre de mots, et m le nombre de documents ; u_{ij} est binaire, il vaut 1 si le mot i apparait dans le document j.

On a discuté la possibilité d’extraire des mots ultra-fréquents. On veut savoir comment les liens entre ces mots sont organisés et regroupés dans leur globalité. Un clustering brutal conduisant à une visualisation directe de tous les liens entre les mots du corpus donnerait une boule relativement sombre sans irrégularité. Pour pallier ce souci, on se contente d’afficher le voisinage de chaque mot. Deux mots sont considérés comme voisins s’ils co-apparaissent dans le même document. On doit imaginer une réduction des données pour améliorer la qualité de la visualisation. L’algorithme présenté dans [Turenne, 2013] propose une réduction de recentrage par la moyenne.

Le calcul de la matrice d’incidence se base sur un simple recentrage des données par rapport à la moyenne des valeurs non nulles : $Y = M.{}^{t}M - \beta.< M >_{l}$, où $<M>_{l}$ est le vecteur moyen en ligne de M et ${}^{t}M$ est la transposée de M. Le paramètre β joue un rôle de régulation pour obtenir plus ou moins de voisins. Si β vaut 1 on recentre les données par rapport à la moyenne des documents dans lesquels ils cooccurrent et on affiche les liens vers ceux qui apparaissent plus que la moyenne. On ne définit pas le nombre de voisins a priori.

L’algorithme d’affichage Drl utilisé pour afficher l’ensemble des liens tend à produire des clusters qui s’opposent visuellement, tandis que l’algorithme de Fruchterman-Reingold tend à produire des clusters circulaires, c’est-à-dire juxtaposés.

Comme hypothèse de visualisation on se fixe à identifier la structure globale en fonction des intervalles de fréquences. Selon le caractère typiquement zipfien de la répartition des mots, un intervalle d’occurrences [f-F] nous semble être un bon critère pour paramétrer au mieux numériquement un espace des mots. Un mot i est retenu si $\le \sum_j u_{ij} \le F$.

En choisissant l'intervalle [2-9] et en regardant le résultat de la visualisation on s'aperçoit que la densité des mots plus fréquents associés entre eux est différente de celle des mots peu fréquents associés entre eux. Les mots plus fréquents modulent la densité des mots peu fréquents qui créent la structure en classes.

Les motifs ont été définis à partir de verbes issus du corpus. L'énumération de tous les verbes nous a conduits à une liste de 291 verbes qui se déclinent, dans le corpus (passé, gérondif..) en une liste étendue de 705 formes distinctes. Dans les différentes extractions de clustering global (figure 4), les points correspondant à cette liste étendue ont été coloriés en rouge. Plusieurs dizaines de mots se retrouvent proches des clusters ce qui indique un choix délibéré d'usage de ces mots dans le contexte de la rédaction des projets.

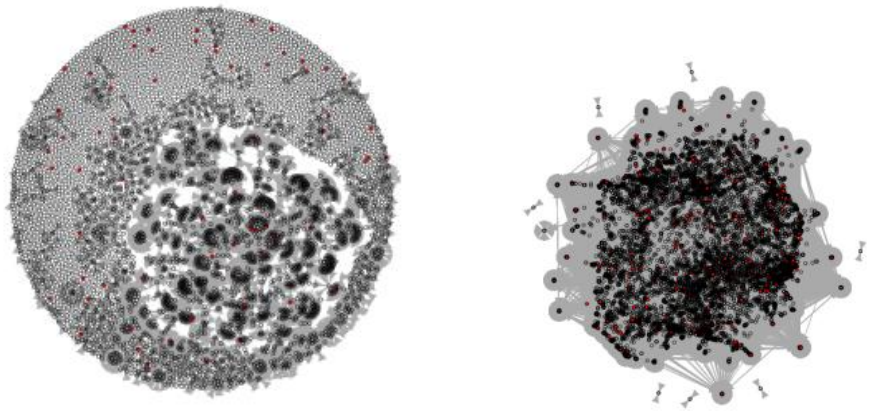


FIG. 4 – Affichage des mots bruts du grand corpus. Intervalle du nombre d'occurences [2-20] ; β=1. Le nombre moyen de voisins est de 32, N=12399 mots (DRL, à droite), β=5, Le nombre moyen de voisins est de 1, N=4985 mots (Fruchterman, à gauche).

4.3.4 Visualisation locale avec des graphes de collocation

Si la visualisation globale permet d'estimer l'intérêt potentiel d'exploiter un ensemble d'éléments lexicaux, une visualisation locale est nécessaire pour procéder à l'intégration des éléments lexicaux dans des règles. Ceci parce que des éléments ne sont pas susceptibles de constituer une règle en tant que tels. Des relations d'équivalence (par exemple groupe de verbes cognitivement équivalents) ou d'attachement syntaxique (par exemple une association nom-verbe) forment des combinaisons lexicales non-ambigues et pertinentes. Pour ce faire, nous avons eu recours au logiciel TermoStat.

TermoStat (2002) est un dépouilleur terminologique en ligne, développé par P. Drouin. En ce qui concerne la présentation des résultats de l'extraction, elle est structurée en cinq fenêtres différentes : Liste des termes, Nuage, Statistiques, Structuration et Bigrammes. La dernière fenêtre de l'interface du logiciel, Bigrammes, liste les collocations binaires Verbe-Nom que le logiciel a repérées dans l'analyse. De même que pour la fenêtre Liste des termes, en cliquant sur les termes affichés dans la fenêtre Bigrammes il est possible d'accéder à une autre fenêtre de Contextes qui affiche les contextes et les concordances. Un lien hypertextuel

en jaune à droite de chaque unité lexicale listée dans la fenêtre Bigrammes permet d'accéder à une autre fenêtre, Décomposition, qui affiche les autres éventuelles collocations dans lesquelles entre l'unité lexicale.
L'extraction terminologique a été limitée aux catégories des noms et des verbes. Suite à cette opération, nous avons pu :

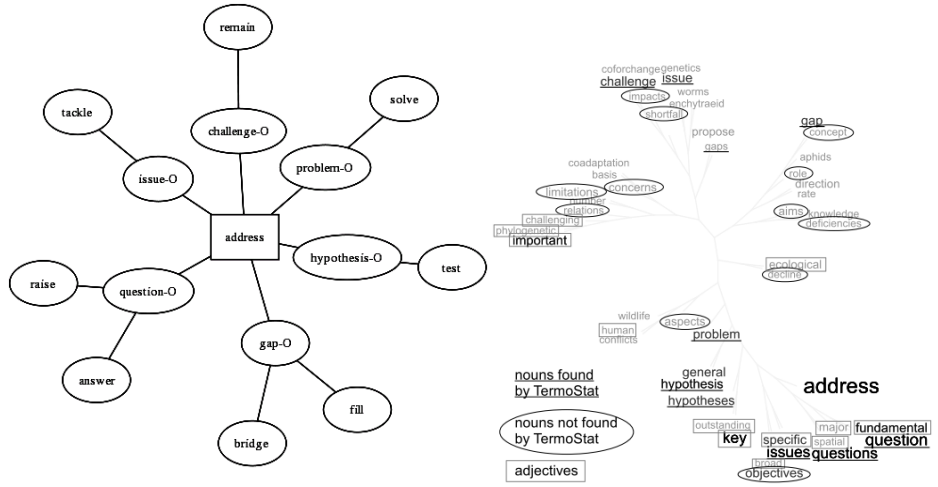
1. Identifier, sur la base de l'extraction nominale, les notions propres à la biodiversité que nous avons citées plus haut ;
2. Classifier les verbes en 6 classes sémantiques – très générales – pour pouvoir ensuite les utiliser dans les motifs : "use verbs" (verbes désignant des méthodes), "goal verbs" (verbe d'objectif, très importants), "technical verbs" (verbes techniques de la biodiversité), "positive verbs" (verbes exprimant une amélioration ou un concept positif), "negative verbs" (verbes exprimant au contraire une péjoration ou un concept négatif), "quantity/evaluation verbs" (verbes exprimant une évaluation quantitative ou un changement de taille) ;
3. Repérer, à partir des bigrammes extraits et des motifs conceptuels fournis dans le logiciel, d'autres collocations verbe-nom fréquentes dans le corpus, afin de réaliser un motif exploitable dans Unitex.

4.3.5 Visualisation locale avec des graphes de similarité

TreeCloud permet de visualiser des informations d'occurrence et de co-occurrence d'un ensemble de mots dans un corpus [Gambette et Véronis, 2009]. Pour cela, il dispose les mots choisis (par défaut, les plus fréquents), dont la taille reflète le nombre d'occurrences, autour d'un arbre qui rapproche au mieux les termes qui apparaissent à proximité dans le texte.

Une telle visualisation peut s'employer sur l'ensemble du corpus afin d'en fournir un résumé visuel. Elle peut aussi être construite à partir des concordances d'un mot d'intérêt. En cela, elle complète les visualisations des bigrammes sous forme de graphes construits par TermoStat. Contrairement à ces derniers, elle ne se limite pas à des catégories grammaticales données mais permet de visualiser l'ensemble des voisins, à gauche, ou à droite, selon les données fournies en entrée, d'un terme d'intérêt, en regroupant ces voisins en communautés de termes sémantiquement proches. Chacun de ces groupes de termes apparaît comme un sous-arbre interprétable par le lecteur de la visualisation [Amstutz et Gambette, 2010].

Nous avons souhaité tester si les bigrammes verbes-noms repérés par TermoStat peuvent être enrichis, c'est-à-dire déterminer si d'autres noms sont fréquemment utilisés comme compléments d'objet des verbes ciblés, et seraient donc pertinents comme éléments de base dans la construction de motifs d'argumentation. Pour cela, nous utilisons les visualisations construites par TreeCloud, en réalisant les nuages arborés des voisinages de mots d'intérêts.



(a) (b)

FIG. 5 – *Le schéma conceptuel TermoStat des bigrammes address-[nom], et le nuage arboré TreeCloud des 50 mots les plus fréquents, hors mots vides, dans les contextes droits de address.*

Pour chacun d'entre eux, nous avons construit :

- le graphe des bigrammes verbe-noms dans TermoStat afin d'identifier les noms le plus souvent utilisés comme complément d'objet du verbe cible ;
- la visualisation en un nuage arboré TreeCloud des mots les plus fréquents (50 au maximum) dans un contexte de k mots à droite du verbe cible, pour k variant de 2 à 4 ;

Nous nous attendons à retrouver les noms extraits par TermoStat dans le nuage arboré construit par TreeCloud, ainsi que d'autres noms non repérés par TermoStat, mais également pertinents puisqu'utilisés comme compléments d'objet du même verbe. Cette hypothèse est vérifiée, comme le montre l'exemple de la figure 5.

Le nombre maximum de mots qui apparaissent dans le nuage arboré a été fixé à 50 de manière à limiter le temps de lecture nécessaire à l'identification de noms pertinents. Nous constatons que cette valeur est toutefois assez grande pour nous permettre de retrouver les noms identifiés par TermoStat.

Pour comparer les résultats obtenus en fonction de la valeur de k, un tableau récapitulatif des résultats est donné par le tableau 3. Il montre les performances les moins bonnes pour k=2 et un léger avantage pour k=3 par rapport à k=4. Notons que le seul cas où un mot repéré par TermoStat n'est pas présent dans le nuage arboré TreeCloud est le nom country pour le verbe develop. Il s'agit là d'une erreur d'étiquetage morphosyntaxique par TermoStat puisque les occurrences de cette association dans le corpus sont exclusivement dues aux noms composés developed countries et developing countries.

Le tableau fait toutefois apparaître une limite des nuages arborés, peu surprenante en raison de notre choix de ne pas effectuer de traitement morphosyntaxique automatisé dans cette approche semi-automatique d'enrichissement : le nombre des termes pertinents ne s'élève qu'à un tiers environ de l'ensemble des termes représentés. Ainsi, une lecture humaine est nécessaire pour distinguer les termes pertinents du bruit constitué par les autres mots. L'organisation sémantique des mots autour de l'arbre facilite toutefois le repérage des termes

pertinents en fournissant pour chacun, à travers l'ensemble de ses voisins dans l'arbre, une sorte de résumé du contexte d'utilisation.

verbe	address	answer	develop	propose
noms compléments d'objet trouvés par TermoStat	6	1	*10*	5
noms Termostat extraits par TreeCloud pour *k*=2	**6**	**1**	**9**	4
noms Termostat extraits par TreeCloud pour *k*=3	**6**	**1**	**9**	**5**
noms Termostat extraits par TreeCloud pour *k*=4	**6**	**1**	**9**	**5**
noms supplémentaires extraits par TreeCloud pour *k*=2	**12**	0	8	9
noms supplémentaires extraits par TreeCloud pour *k*=3	**12**	**3**	**9**	**12**
noms supplémentaires extraits par TreeCloud pour *k*=4	**12**	**3**	**9**	9

TAB. 3 – *Résultat de la méthodologie d'extraction de noms apparaissant comme complément d'objet direct autour de verbes d'intérêt.*

De plus, ces mots supplémentaires peuvent avoir un intérêt en tant que tels : plusieurs d'entre eux sont des adjectifs, notamment des adjectifs à connotation positive (broad, challenging, fundamental, important, key, major, outstanding), qui apparaissent au coeur de la démarche argumentative exprimée par les bigrammes verbes-noms analysés.

4.4 Construction des motifs

4.4.1 À partir de l'extraction terminologique

Si TermoStat est un logiciel conçu en vue du traitement de corpus spécialisés, Unitex [Paumier, 2002] est une collection de programmes pour le traitement de textes en langues naturelles sur la base de ressources lexicales. Les motifs sont des transducteurs représentables sous forme de graphes d'automates finis [M. Gross, 1997]. Unitex possède des potentialités très intéressantes pour la recherche de contextes. Ces recherches peuvent être menées à partir du programme Locate Pattern, qui sert de "moteur de recherche" sur le texte analysé. Le programme Locate Pattern est, en outre, étroitement lié à l'application de grammaires locales. L'utilisateur, en fait, peut choisir de mener sa recherche par expression régulière (unité lexicale, catégorie grammaticale, patron syntaxique) ou par application d'une grammaire locale. Il est possible d'afficher toutes les occurrences dans le texte visées par la recherche ou de limiter la recherche à un nombre restreint (200 formes par défaut). Pour chaque concordance recherchée, un fichier HTML nommé "concord" est produit dans le dossier .snt du texte traité et il est consultable indépendamment d'Unitex. Les occurrences s'affichent sous formes de liens hypertextuels, qui renvoient au contexte d'utilisation d'une forme.

Dans Unitex, il est possible de réaliser des motifs à partir du menu Graph et de pouvoir ensuite les appliquer depuis la fonction Locate Pattern. L'annexe 1 donne un exemple d'application d'un motif.

Comme nous l'avons dit plus haut, nous avons utilisé une partie des résultats de l'extraction terminologique dans TermoStat pour la création de nos motifs. Dans ce même but, nous avons exploité les résultats de l'extraction des mots ayant une fréquence supérieure

à 100 dans R. A l'heure actuelle, notre base de données compte 30 motifs, divisés en 5 groupes distincts selon le type d'information recherché. Nous avons donc : un groupe de motifs généraux, un groupe de motifs de possibilité, un groupe de motifs détectant des noms cibles, un groupe de motifs détectant des verbes cibles et un dernier groupe de motifs visant la recherche de connecteurs. On peut comprendre que, à l'égard des objectifs que nous nous sommes fixés, les motifs généraux ont une plus grande importance par rapport aux autres, bien que les autres groupes de motifs aussi contribuent à la recherche d'informations liées à l'argumentation.

Les 10 motifs du groupe de motifs généraux sont les plus complexes. Les motifs repérés par ces motifs expriment surtout des buts que les acteurs des projets se fixent ou bien des méthodes mises en place pour atteindre ces buts. Pour la réalisation de ces motifs, nous avons accordé une importance particulière aux emplois verbaux, notamment, à l'emploi du futur, de l'infinitif et du gérondif. De même, nous avons privilégié les verbes appartenant aux classes sémantiques identifiées après l'extraction dans TermoStat. La figure 6 montre l'automate argument.fst2, qui reconnaît les phrases qui contiennent un nom désignant un objectif (goal noun) en position sujet suivi d'une phrase infinitive dans laquelle apparaît soit un verbe de but (goal verb), soit un verbe exprimant des méthodes utilisées dans le projet (use verb), soit un verbe à valeur positive (positive verb), soit un verbe d'évaluation (quantity/evaluation verb).

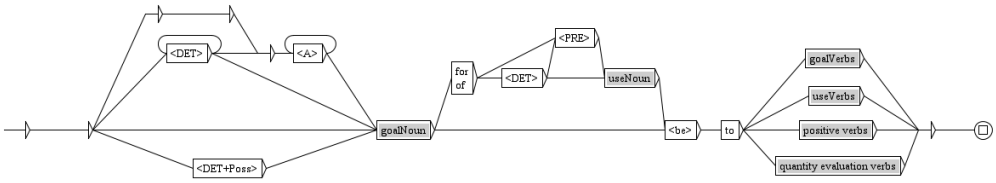


FIG. 6 – *Le graphe argument.fst2.*

En guise d'exemples de résultats produits par l'application de ce graphe, nous citons les phrases suivantes :

- Main purposes are to collect… ;
- Our global objective is to relate… ;
- The aim of the present project is to calculate…

Un autre motif appartenant à la catégorie des motifs généraux est le motif argument3 (figure 7). Ce graphe a été conçu pour la reconnaissance des phrases contenant le futur avec will.

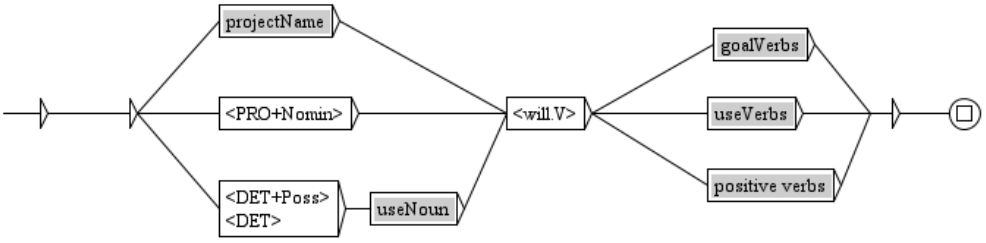


FIG. 7 – *Le graphe argument3.*

Quelques exemples :

- My analyses will produce… ;
- Both experiments will focus… ;
- This research project will undertake…

Avec le motif argument.grf et avec deux autres motifs (argument2.grf et argument4.grf), ces deux motifs constituent un motif à très large couverture, que nous avons appelée argumentation.grf.

Dans le groupe des motifs généraux, il y a aussi le motif likely.grf (figure 8), qui recherche des phrases construites autour de l'adjectif *likely*, très intéressant pour le repérage des phrases qui servent d'ouverture sur le sujet du projet. *Likely* utilisé en anglais dès qu'on n'est pas sûr de quelque chose soit dans le cadre du contexte du projet (extreme events are likely to increase in the future….) soit pour prendre des précautions, en termes d'assertions scientifiques (eg. 'environmental variables likely driving biodiversity') ou en termes de délivrables du projet 'the project will likely help farmers to develop sustainable practices… ', 'The project is also likely to deliver'.

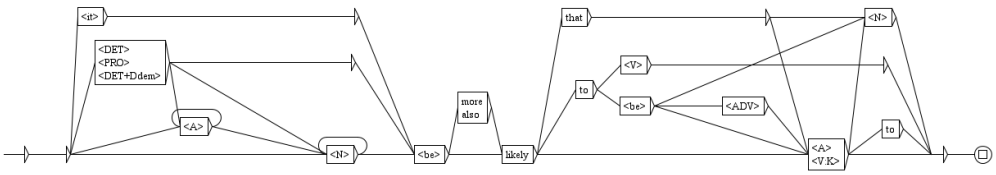


FIG. 8 – *Le motif likely.grf.*

Ainsi, le motif reconnaît les phrases :

- Climate change is likely to increase…;
- Energy expenditure is likely to correlate…;
- The project is also likely to deliver…

4.4.2 A partir de patrons inspirés de la théorie de l'argumentation

Au-delà des grammaires construites sur les résultats de l'extraction terminologique dans TermoStat, nous avons également essayé d'exploiter certains patrons syntaxiques ou morphologiques traités dans quelques-uns des manuels sur l'argumentation passés en revue.

Nous nous sommes inspirés principalement de trois modèles théoriques : la performativité des actes locutoires [Austin], l'ADL (Argumentation Dans la Langue) de [Ducrot et Anscombre] et l'approche pragma-dialectique au discours argumentatif [van Eemeren et al.].

Par exemple, les 10 automates du groupe "argumentation" ont été conçus avec une attention particulière à la performativité des actes locutoires, théorisée par [Austin 1955].

Le travail mené par Ducrot et Anscombre [1980, 1983] sur le rôle des connecteurs en discours nous a inspiré pour la réalisation des automates du groupe « connecteurs », en adaptant, bien sûr, cette approche à la langue anglaise. L'automate "step" peut également être inclus dans ce groupe, car il reconnaît les phrases qui servent à agencer les arguments, comme par exemple : Our first step will be to…

Pour en venir à l'influence de l'École d'Amsterdam [van Eemeren et al.], nos avons repris et enrichi certains des motifs syntaxiques proposés pour exemplifier les trois types principaux d'argumentation identifiés par ces chercheurs et que nous avons cités plus haut. Pour chaque type d'argumentation, les auteurs listent des phrases simples contenant un verbe qui exprime la relation existant deux éléments X et Y, qui peuvent être un point de vue et un argument. En guise d'exemple, quelques-unes des phrases simples pour fournir des indications de relations symptomatiques unidirectionnelles :

X implies Y…

X indicates Y…

X means Y…

C'est précisément le type de relation exprimée par nos automates contenant un « use verb ».

Comme nous l'avons répété plusieurs fois, nous ne sommes qu'au début de notre travail, qui semblent avoir des débouchées ultérieures. Donc, il va sans dire que la base des automates est provisoire et a de fortes possibilités d'être enrichie dans le temps et grâce à une étude plus approfondie de l'immense littérature sur l'argumentation et l'extraction d'information.

4.5 Caractérisation du genre "résumé de projet" par l'application des motifs dans Unitex

4.5.1 Énumération des résumés faisant apparaître les motifs d'argumentation

Afin d'évaluer le nombre de résumés de projets faisant apparaître chacune des 30 motifs d'argumentation construites dans Unitex, un script Python a été préparé. Il découpe l'ensemble du corpus en un fichier par résumé, puis appelle Unitex pour appliquer chacun des motifs, et stocke l'ensemble des occurrences détectées pour le motif utilisé, avec un contexte d'une quarantaine de caractères de part et d'autre de la forme repérée. Ce sont ces fichiers d'occurrences qui sont ensuite parcourus pour évaluer, pour chaque résumé, le nombre d'occurrences de chaque motif d'argumentation. La figure 9 affiche, selon un code couleur pour chaque motif, les éléments textuels correspondant au motif présent dans le résumé.

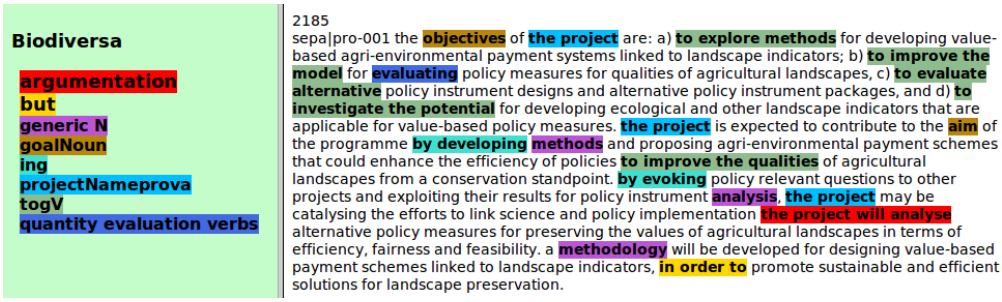


FIG. 9 – *Mise en surbrillance des éléments lexicaux des motifs détectés dans un résumé en particulier. Huit motifs sont présents.*

Ces données sont alors utilisées pour calculer le nombre d'occurrences de chaque motif dans l'ensemble du corpus, mais aussi le pourcentage de résumés du corpus qui utilise chaque motif. Enfin, un script de rééchantillonnage aléatoire permet de calculer ces deux paramètres pour des sélections aléatoires de 1000 projets : une moyenne sur 1000 sélections aléatoires peut alors être calculée, ainsi qu'un écart-type, afin d'évaluer la stabilité des résultats obtenus avec chaque motif.

4.5.2 Efficacité des motifs

Le tableau 4 résume le nombre de fois que chaque motif trouve une séquence avec succès. D'autre part il est calculé le pourcentage de documents du contenant une séquence correspondant à un motif (taux de couverture). Un protocole de calcul d'erreur a été mis en place pour estimer la variation du pourcentage de couverture. Le taux de couverture est aussi calculé pour chaque motif et pour un échantillon du corpus d'une même année de publication. L'ensemble des valeurs du taux de couverture traduit l'évolution de l'emploi des motifs d'argumentation (figure 10). On constate que le taux augmente pour tous les motifs sauf un (argument2) entre les périodes [2000-2005] et [2009-2011]. Il augmente de plus de 10 points de pourcentage pour les 15 motifs de la figure 10. Compte tenu de l'écart-type, cette différence est significative. Elle correspond probablement à une professionnalisation des scientifiques dans leur manière de répondre à des appels à projets dont le format est assez codifié, voire aussi à l'arrivée de nouveaux acteurs qui emploient plus facilement ces motifs d'argumentation.

Un taux élevé de résumés (87,69%) indique les acteurs des projets, qui proposent des solutions à des questions. Parmi les motifs développés à partir de l'étude des connecteurs dans les corpus, celles repérant les expressions de la cause, du but et de la conséquence ont les taux de couverture les plus élevés. A remarquer également le taux de couverture du motif « ajouter », qui identifie les connecteurs permettant une énumération des arguments des auteurs des résumés.

Motif	Nombre d'occurrences trouvées parmi les 4655	Pourcentage des résumés avec au moins une occurrence du motif	Ecart-type des pourcentages pour 1000 sélections aléa-

Définition semi-automatique de motifs d'argumentation

	résumés	d'argumentation	toires de 1000 résumés
Motifs généraux			
Argumentation (motif contenant les 4 motifs ci-dessous)	1626	47,43%	1,40
Argument	4914	14,09%	0,97
Argument2	3	0,06%	0,07
Argument3	3060	37,64%	1,38
Argument4	26	0,54%	0,21
Ing	1979	28,85%	1,26
Shall	567	8,87%	0,79
Step	160	2,84%	0,46
Likely	268	5,07%	0,61
ToV	6876	69,69%	1,22
Motifs de possibilité			
Goal N can	16	0,34%	0,16
Positive N can	41	0,88%	0,27
Generic N can	151	3,05%	0,48
Types de verbes			
Negative verbs	3538	39,38%	1,33
Quantity evaluation verbs	7236	64,04%	1,33
Types de N			
Positive N	5300	57,47%	1,38
Goal N	6172	64,34%	1,39
Project name rev	15415	87,69%	0,96
Connecteurs			
Début	261	5,18%	0,61
But	1139	19,21%	1,12
Concession	3097	42,53%	1,37
Agencement	1626	25,22%	1,23
Cause (sans as)	2685	35,64%	1,31
As	7405	66,53%	1,29
Ajouter	4914	53,83%	1,35
Appositive	2237	29,6%	1,27
Comparaison	331	6,1%	0,66
Conséquence	3431	42,47%	1,36
Hypothèse	2572	29,77%	1,29
Introduire	168	3,33%	0,50

TAB. 4 – *Nombre d'occurrences du corpus pour chaque motif (avec le pourcentage de couverture sur les documents et l'erreur).*

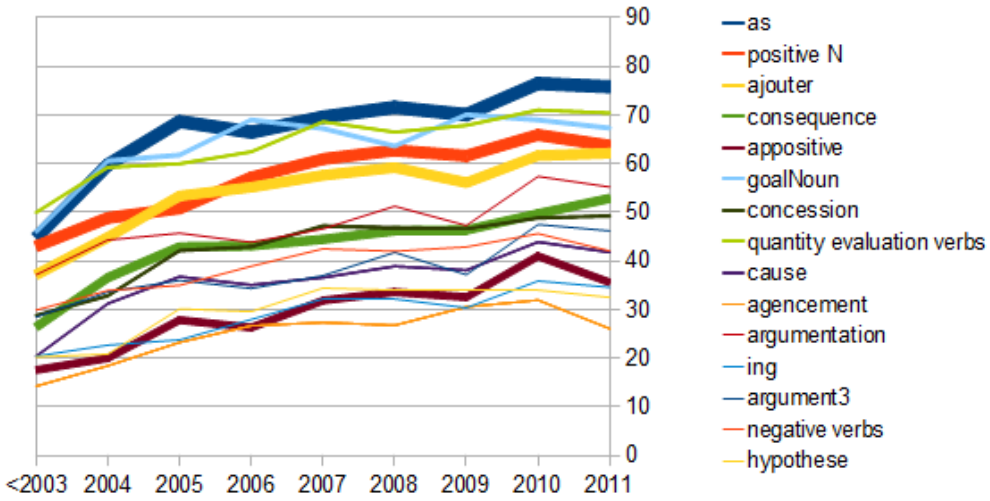


FIG. 10 – *Taux de couverture, pour chaque sous-corpus défini par année de publication, des motifs d'argumentation montrant la plus forte hausse entre 2000-2006 et 2007-2013.*

La table 5 montre le taux d'erreur (le nombre de documents oubliés par un motif) que l'on calcule comme suit $\frac{VP-(PM-FP)}{VP}$ où VP est le nombre de vrais positifs évalué à la main auquel on soustrait le nombre corrigé (par les faux positifs) du nombre de positifs par l'application automatique du motif. Certaines formulations n'ont pas été intégrées par le motif ce qui explique cet oubli comme : « demographic effects of escaped farmed atlantic salmon on wild salmon populations » où l'on voit l'expression significative « effects…on… » est ignorée par le motif actuel. La couverture de 42% (motif conséquence) et 12% (motif hypothèse) pour l'échantillon de 50 documents est respectivement proche des 42% et éloignée des 29% sur l'ensemble du corpus ce qui nous laisse à penser que l'erreur peut varier d'un motif à l'autre. L'erreur n'est pas catastrophique et il semble raisonnable au vue des expressions oubliées d'améliorer l'erreur.

motif	**PM - Positifs (motif)**	**FP - Faux Positifs (motif)**	**VP - Vrais Positifs**	**taux d'erreur (faux négatifs)**	**Couverture (VP)**
conséquence	23	2	38	*44,7%*	42%
hypothèse	17	11	9	*33,3%*	12%

TAB. 5 – *Taux d'erreur, faux positifs ou oubli des vrais positifs, pour deux motifs en particulier, le motif conséquence et le motif hypothèse.*

5 Perspectives

Une perspective est l'exploitation multi-dimensionnelle des attributs d'argumentation obtenus par l'application des motifs.

La plateforme CorText permet d'exploiter les informations géographiques sur les labos ou universités. Cette information pourrait être intéressante pour contextualiser les thèmes et les liens territoriaux entre universités.

Une autre perspective est l'exploitation des informations sur les thématiques abordées par les projets financés référencés dans la base de données BiodivERsA. L'application d'outils de statistique textuelle (calcul et/ou visualisation), après prétraitement impliquant les motifs pour le pré-traitement du texte (mots composés), permettra de caractériser les grands thèmes de recherche abordés par les projets du corpus. On voit ici tout le potentiel d'une telle analyse pour caractériser des évolutions temporelles dans les thèmes privilégiés par la recherche sur la biodiversité en Europe, ainsi que pour comparer les thèmes de recherche majeurs soutenus par différents programmes ou par différents pays européens.

Finalement, le critère objectif que nous avons utilisé pour rendre compte de la pertinence d'un motif d'argumentation est un pourcentage de couverture sur l'ensemble du corpus. Il ne fait pas état de la combinaison d'arguments à savoir si les projets qui utilisent un argument (par exemple « objectif ») font aussi état d'un autre argument (par exemple « hypothèse »). Une autre perspective de notre étude devrait aborder l'extraction des contextes droits des motifs comme « our hypothesis is that » qui renvoie des contextes comme « polymorphic Y chromosome differs in the number of binding sites » et « there is a critical level of genetic diversity for stress responses ». Nous prévoyons d'étudier comment extraire des entités nommées à partir des contextes droits.

Nous croyons également que le test de ces motifs pourrait révéler que ces motifs d'argumentation ne sont pas une prérogative absolue du domaine traité, mais qu'ils constituent des formules typiques des écrits scientifiques qui ont le but de convaincre. Toutefois, pour valider cette hypothèse, il serait nécessaire d'appliquer ces motifs à d'autres corpus du même genre textuel.

6 Conclusion

Notre hypothèse de départ a été de détecter et analyser des motifs d'argumentation utilisés par les scientifiques à travers l'étude des projets du domaine de la biodiversité. Cela nous a amenés à considérer des aspects sémantiques d'extraction d'information non standard tels que la définition de motifs d'argumentation. Ces motifs indiquent la présence d'un argument sans l'extraire de manière explicite ; il s'agit davantage de marqueurs que d'arguments en tant que tels. Plusieurs outils d'extraction d'items lexicaux : TermoStat, TreeCloud, TM(R), GlobalPlot(R), nous ont aidés à définir manuellement 30 motifs d'argumentation. Ces motifs ont été exploités par un outil d'application de motifs (Unitex) et un script Python d'analyse du taux de couverture des motifs. L'utilisation de ces six outils révèle en premier lieu une certaine difficulté d'extraire des associations locales pertinentes. Néanmoins notre travail montre, à travers des éléments d'évaluation en termes de taux de couverture sur l'ensemble des projets, que certains motifs sont très productifs sur l'ensemble du corpus. D'autre part un point de vue cinétique sur l'usage des motifs montre une intensification notable entre les premières et les dernières années. Il semblerait que, dans un contexte où les appels à projets

compétitifs représentent une source de financement de plus en plus important, les chercheurs du domaine de la biodiversité s'adaptent à ce mode de financement et professionnalisent la façon dont ils répondent, en utilisant de façon de plus en plus systématique des motifs d'argumentation résumant clairement leurs objectifs et type de connaissance attendue. Cela laisse à penser que la rédaction est construite en fonction d'arguments précis pour faciliter la lecture et la recommandation d'un projet. Si beaucoup reste encore à faire pour mieux apprécier l'impact de l'argumentation sur la triade institutions-projets-domaine, nous pensons que l'approche décrite dans cette étude pourrait convenir au traitement d'autres genres de documents et d'autres domaines.

Remerciements

Ce travail utilise les données référencées dans la base de données BiodivERsA au 15 juin 2013. L'ensemble des partenaires de BiodivERsA sont vivement remerciés pour leur investissement important dans l'élaboration de cette base de données. BiodivERsA est un ERA-NET soutenu par le 7ème Programme Cadre de la Commission Européenne. Nous remercions aussi le support du projet PAN-Bioptique (01-2012-06-2013), "Les nouvelles institutions de la biodiversité : Inventorier, numériser, expertiser la nature", financé par l'Agence nationale de la recherche (ANR programme Sciences, technologies et savoirs en société : enjeux actuels, questions historiques).

Références

Adelman Leonard , Paul E. Lehner, B. A. Cheikes, M. F. Taylor: An Empirical Evaluation of Structured Argumentation Using the Toulmin Argument Formalism. IEEE Transactions on Systems, Man, and Cybernetics, Part A (TSMC) 37(3):340-347 (2007)

Amgoud L. et Prade H. (2012) Can AI Models Capture Natural Language Argumentation? International Journal of Cognitive Informatics and Natural Intelligence 6(3), 14 pages.

Amstutz D. et Gambette P. (2010) Utilisation de la visualisation en nuage arboré pour l'analyse littéraire,Proceedings of the 10th International Conference on statistical analysis of textual data (JADT'10), Statistical Analysis of Textual Data, p. 227-238.

Anscombre J.C. et Ducrot O., L'argumentation dans la langue, Mardaga, Coll. "Philosophie et langage", 1983.

Austin, J.L. (1955), How to do things with words, Harvard University Press.

Baldridge J, Asher N et Hunter J. Annotation for and Robust Parsing of Discourse Structure on Unrestricted Texts", Zeitschrift fur Sprachwissenschaft , 26 (2007), 213-239.

Baldridge J, Asher N et Hunter J. Annotation for and Robust Parsing of Discourse Structure on Unrestricted Texts", Zeitschrift fur Sprachwissenschaft , 26 (2007), 213-239

Barré, R., 2001. Sense and nonsense of ST productivity indicators. *Science and Public Policy*, 28, 259-266.

Beaver, D., 2001. Reflections on Scientific Collaboration (and its study): Past, Present, and Future. *Scientometrics*, 52(3), 377, 365.

Bench-Capon Trevor J. M. : The Long and Winding Road: Forty Years of Argumentation. COMMA 2012:3-10

Besnard P. , Hunter A.: Elements of Argumentation. MIT Press 2008

Bonaccorsi, A., 2008. Search Regimes and the Industrial Dynamics of Science. *Minerva*, 46(3), 285-315.

Boullier D. et Lohard A (2012) Opinion mining et sentiment analysis. Open Press, Marseille.

Bourret P., Mogoutov A., Julian-Reynier C., et Cambrosio A., (2006). A New Clinical Collective for French Cancer Genetics A Heterogeneous Mapping Analysis, Science, Technology, & Human Values, 31 (4): 431-464.

Bozeman, B. et Rogers, J.D., 2002. A churn model of scientific knowledge value: Internet researchers as a knowledge value collective. *Research Policy*, 31(5), 769-794.

Braud Ch. et Denis P., Identification automatique des relations discursives "implicites" à partir de données annotées et de corpus bruts, in Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013). Volume 1: TALN-RECITAL. pp.104-117.

Breton Ph. et Gauthier G. 2000. Histoire des théories de l'argumentation. Paris : La Découverte, coll. « Repères ».

Budzynska Katarzyna , Chris Reed: Speech Acts of Argumentation: Inference Anchors and Peripheral Cues in Dialogue. Computational Models of Natural Argument 2011

Cabrio E et Villata S. Natural Language Arguments: A Combined Approach, in Proceedings of 20th biennial European Conference on Artificial Intelligence, ECAI 2012, Montpellier, France.

Callon, M., J. Law, A. Rip (1986), Mapping the Dynamics of Science and Technology. London: The MacMillan Press Ltd.

Cambrosio A., Keating P., Mercier S., Lewisonc G., et Mogoutov A., (2006). Mapping the emergence and development of translational cancer research, European journal of cancer, 24: 3140-3148

Cambrosio A., Keating P., Mogoutov A. (2004). Mapping collaborative work and innovation in biomedicine: a computer assisted analysis of antibody reagent workshops, Social Studies of Science, 34 (3): 325-364.

Caminada Martin , Leila Amgoud: On the evaluation of argumentation formalisms. Artif. Intell. (AI) 171(5-6):286-310 (2007)

Chaveriat C., Ghitalla F., Pelegrin F., Fadil F. & Le Roux X. (2011). La base de données nationale des acteurs, structures et projets de recherche sur la biodiversité: présentation et analyse du paysage de la recherche. Rapport FRB, Série Expertise et synthèse, 36 pages.

Chesnevar C, Maguitman A.G. et Gonzalez M.P. (2009) Empowering Recommendation Technologies Through Argumentation, Argumentation in Artificial Intelligence. ISBN 978-0-387-98196-3. Springer-Verlag US, p. 403.

Cover TM et Hart PE (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13 (1): 21-27.

Csardi G et Nepusz T (2006). The igraph software package for complex network research, InterJournal, Complex Systems 1695.

Culioli Antoine, Pour une linguistique de l'énonciation : formalisation et opérations de repérage, tome 2, Paris, Ophrys, 1999.

de Jonge Emmanuel , « Pertinence de l'utilisation du modèle de Toulmin dans l'analyse de corpus », Argumentation et Analyse du Discours [En ligne], 1 | 2008, mis en ligne le 18 septembre 2008

Doucet A. et Ahonen-Myka H. Fast extraction of discontiguous sequences in text: a new approach based on maximal frequent sequences in Proceedings of IS-LTC 2006, Information Society - Language Technologies Conference, Ljubljana, Slovenia, October 9-14, 2006, p. 186-191.

Ducrot Oswald et Jean-Claude Anscombre, L'argumentation dans la langue, Mardaga, 1983

Eemeren, F.H. van et Grootendorst, R. (1992) Argumentation, Communication, and Fallacies. Hillsdale, NJ: Lawrence Erlbaum.

Eggermont H., Le Roux X., Heughebaert A., Balian E. & BiodivERsA partners (2013). The BiodivERsA database: Analysis of the competitive funding landscape for research on biodiversity and ecosystem services in Europe. BiodivERsA report, 33 pp.

Enguehard C., Daille B., et Morin E. Tools for terminology processing. In Proceedings of The Indo-European Conference on Multilingual Communication Technologies (IEMCT), pages 218-229, 2002.

Feinerer I., Hornik K. et Meyer D (2008). Text mining infrastructure in R. Journal of Statistical Software, 25(5):1-54.

Feng V. W. et Hirst G. 2012. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-2012)*, Jeju, Korea.

Fréard Dominique , Alexandre Denis, Françoise Détienne, Michael Baker, Matthieu Quignard, Flore Barcellini: The role of argumentation in online epistemic communities: the anatomy of a conflict in Wikipedia. ECCE 2010:91-98

Freeman, C. et Soete, L., 2009. Developing science, technology and innovation indicators: What we can learn from the past. *Research Policy*, 38(4), 583-589.

Fruchterman T. et Reingold E. (1991), Graph Drawing by Force-Directed Placement. Software – Practice & Experience (Wiley) 21 (11): 1129–1164.

Gambette P. et Véronis J. (2009), Visualising a Text with a Tree Cloud. Proceedings of the International Federation of Classification Societies 2009 Conference (IFCS'09), Studies in Classification, Data Analysis, and Knowledge Organization 40, p. 561-570.

Goffman E. Forms of Talk, Philadelphia: University of Pennsylvania Press (1981)

Gómez Sergio Alejandro , Carlos Iván Chesñevar, Guillermo Ricardo Simari: ONTOarg: A decision support framework for ontology integration based on argumentation. Expert Syst. Appl. (ESWA) 40(5):1858-1870 (2013)

Grossmann F. « L'Auteur scientifique », Revue d'anthropologie des connaissances 3/2010 (Vol 4, n° 3), p. 410-426. URL : www.cairn.info/revue-anthropologie-des-connaissances-2010-3-page-410.htm.

Guston, D.H., 2001. Boundary Organizations in Environmental Policy and Science: An Introduction. *Science, Technology & Human Values*, 26(4), 399-408.

Hartley, J. & Betts, L. (2008). Revising and polishing a structured abstract: Is it worth the time and effort? Journal of the American Society for Information Science and Technology, 59, 12, 1870-1877.

Hatzivassiloglou V. et McKeown K. (1997) Predicting the semantic orientation of adjectives, Proceedings of the 8th conference on European chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA, p. 174-181.

Heimeriks, G., Hörlesberger, M. et Van Den Besselaar, P., 2003. Mapping communication and collaboration in heterogeneous research networks. *Scientometrics*, 58(2), 391-413.

Heras Stella , Katie Atkinson, Vicente J. Botti, Floriana Grasso, Vicente Julián, Peter McBurney: Research opportunities for argumentation in social networks. Artif. Intell. Rev. (AIR) 39(1):39-62 (2013)

Hernandez N. et Grau B., 2003, Extraction et typage de termes significatifs pour la description de textes, actes de ISKO, Grenoble.

Ivin, A.A. The theory of argumentation. M.*:* Gardariki, 2000

Jefferys B., Kelley L., Sergot M., Fox J et Sternberg M. (2006) Capturing expert knowledge with argumentation: a case study in bioinformatics. Bioinformatics 22(8):924-33

Kaci Souhila : Refined Preference-based Argumentation Frameworks. COMMA 2010:299-310

Katz, S. et Martin, B., 1997. What is research collaboration? *Research Policy*, 26(1), 18, 1.

Lazega Emmanuel (2011), "Pertinence et structure", Revue Suisse de Sociologie, 37:127-149).

Lebart L., Salem A. et Berry L., Exploring Textual Data, Kluwer Academic Publishers, Boston, 1998.

Lent B., R. Agrawal et R. Srikant, "Discovering Trends in Text Databases", KDD 1997.

Liu B, Hu M.Q. et Cheng J.S. (2005) Opinion Observer: Analyzing and Comparing Opinions on the Web, Proceedings of the 14th international World Wide Web conference (WWW-2005).

Lu Jingyan , Ming Ming Chiu, Nancy WaiYing Law: Collaborative argumentation and justifications: A statistical discourse analysis of online discussions. Computers in Human Behavior (CHB) 27(2):946-955 (2011)

Lucio-Arias, D. et Leydesdorff, L., 2007. Knowledge emergence in scientific communication: from "fullerenes" to "nanotubes". Scientometrics, 70(3), 603-632.

Mann, William C. and Sandra A. Thompson. (1988). Rhetorical Structure Theory: Toward a functionaltheory of text organization. Text, 8(3), 243-281.

Martin S., Brown W., Klavans R. et Boyack K. (2011) OpenOrd: An Open-Source Toolbox for Large Graph Layout. Proceedings of the Visualization and Data Analysis Conference, 2011. Published in SPIE-IS&T. 7868. p. 786806-1-11.

Maurel S., Curtoni P. et Dini L. (2008) L'analyse des sentiments dans les forums, Atelier Fouille des Données d'Opinions (FODOP 08).

McBurney Peter , Simon Parsons: Risk Agoras: Dialectical Argumentation for Scientific Reasoning. UAI 2000:371-379

Mei, Q. et Zhai, C. (2005). Discovering evolutionary theme patterns from text – an exploration of temporal text mining. Proceedings of the 2005 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (KDD'05), Chicago, Illinois, 198-207.

Ontañón Santiago , Enric Plaza: Empirical Argumentation: Integrating Induction and Argumentation in MAS. ArgMAS 2010:49-67

Pak A. et Paroubek P. (2010) Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, European Language Resources Association ELRA.

Perelman, Chaïm et Olbrechts-Tyteca, Lucie. 1988 [1958]. Traité de l'argumentation (Bruxelles, Editions de l'Université Libre de Bruxelles)

Pontille D., 2004, La signature scientifique. Une sociologie pragmatique de l'attribution , Paris, CNRS Editions.

Powell et al., 2005 W.W. Powell, D.R. White, K.W. Koput et J. Owen-Smith, Network dynamics and field evolution: the growth of interorganizational collaboration in the life sciences, American Journal of Sociology, 110 (2005), pp. 901–975.

Prieto José-Luis (1975), Pertinence et pratique: essai de sémiologie, Paris: Editions de Minuit.

R Core Team (2013) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org.

Ricci F, Rokach L, Shapira B et Kantor P (2011) Recommender Systems Handbook. Springer ISBN 978-0-387-85819-7

Saad Missen M.M., Boughanem M. et Cabanac G. (2012) Opinion mining: reviewed from word to document level, Social Network Analysis and Mining, Vienna, Austria, Springer-Verlag Springer.

Searle, John R. 1972 [1969]. Les actes de langage. Essai de philosophie du langage (Paris : Herman)

Smadja F. et McKeown K., "Automatically Extracting and Representing Collocations for Language Generation", Association for Computational Linguistics Conference (ACL), Pittsburgh, United States, 1990.

Song M et Kim S.Y Detecting the knowledge structure of bioinformatics by mining full-text collections, Scientometrics (2013) 96:183–201.

Teufel Simone , Jean Carletta, Marc Moens: An annotation scheme for discourse-level argumentation in research articles. EACL 1999:110-117

Toni Francesca : Assumption-Based Argumentation for Epistemic and Practical Reasoning. Computable Models of the Law, Languages, Dialogues, Games, Ontologies 2008:185-202

Toniolo Alice , Timothy J. Norman, Katia P. Sycara: An Empirical Study of Argumentation Schemes for Deliberative Dialogue. ECAI 2012:756-761

Toulmin, Stephen. 1993 [1958]. Les usages de l'argumentation, trad. P. de Brabanter (Paris : PUF)

Turenne N. (2013) Clustering and Relational Ambiguity: from Text Data to Natural Data. Submitted.

Turenne N. (2013) Knowledge Needs and Information Extraction: Towards an Artificial Consciousness, ISBN: 978-1-84821-515-3, Hardcover 288 pages, Wiley-ISTE.

Turenne N. et Barbier M. BELUGA : un outil pour l'analyse dynamique des connaissances de la littérature scientifique d'un domaine. Première application au cas des maladies à prions, In proceedings of Extraction et Gestion de Connaissances, Hébrail G. and Lebart L. eds., Clermont-Ferrand, 2004, France.

Turney P. et Littman M. (2003) Measuring praise and criticism: inference of semantic orientation from association, ACM TOIS, 21(4), p. 315-346.

Tutin A et Grossmann F. (éd.), L'écrit scientifique : du lexique au discours. Autour de Scientext. Rennes: Presses Universitaires de Rennes. 2014.

Walton Douglas : Using Argumentation Schemes for Argument Extraction: A Bottom-Up Method. IJCINI 6(3):33-61 (2012)

Webber B. et Prasad R. Discourse Structure: Swings and Roundabouts, in Behrens & Fabricius-Hansen (eds.) Structuring information in discourse: the explicit/implicit dimension, Oslo Studies in Language 1(1), 2009. 171-190.

Wyner A., Schneider J., Atkinson K. et Bench-Capon T. (2012) Semi-Automated Argumentative Analysis of Online Product Reviews. Fourth International Conference on Computational Models of Argument (COMMA 2012). September 10-12, 2012. Vienna, Austria

Zipf G.K. (1935) The psychology of Language, an Introduction to Dynamic Philology. Houghton-Mifflin, Boston.

Zitt, M. et Bassecoulard, E., 2008. Challenges for scientometric indicators: data demining, knowledge-flow measurements and diversity issues. Ethics in Science and Environmental Politics, 8, 60, 49.

Summary

.

ANNEXE 1

L'exemple ci-dessous déroule la détection d'une expression lexicale grâce à un motif (représenté sous forme d'un graphe d'états et de transitions. L'expression « our long-term study » est parcourue séquentiellement. La figure 11 montre l'initialisation du motif avant le parcours du mot « our ».

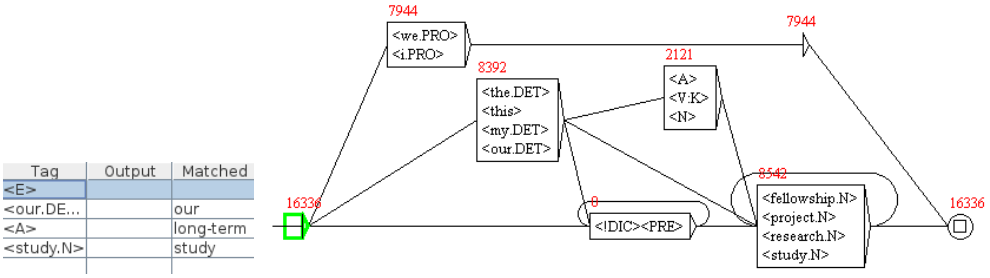


FIG. 11 – *Mot courant, début de l'expression, de la séquence parcourue (à gauche) et nœud (16336) du motif validé correspondant en vert (à droite).*

Le nœud 16336 du motif s'allume. La figure 12 montre le passage au premier mot « our ». A ce niveau une transition du nœud 16336 au nœud 8392 est validée. La figure 13 montre le passage au mot suivant, « long-term », et le motif valide la transition du nœud 8392 au nœud 2121 puisque « long-term » est un adjectif.

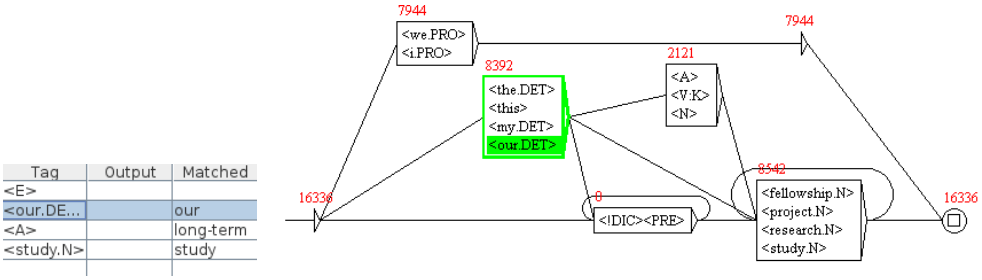


FIG. 12 – *Mot courant, « our », de la séquence parcourue (à gauche) et nœud (8392) du motif validé correspondant en vert (à droite).*

La figure 14 montre le passage au mot suivant « study », par conséquent le motif valide la transition du nœud 2121 au nœud 8542 puisque ce mot fait partie de la liste qui définit l'état du noeud. Finalement il n'y a plus de mot, et la figure 15 montre que le nœud terminal est atteint ce qui valide une transition du nœud 8542 au nœud 16336. Comme la séquence « our long-term study » génère un chemin complet dans le motif, cette expression est détectée comme expression lexicale annotée par ce motif.

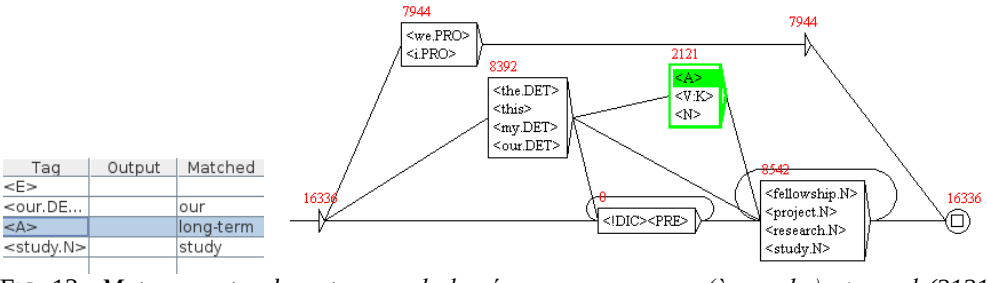


FIG. 13 –*Mot courant, « long-term », de la séquence parcourue (à gauche) et nœud (2121) du motif validé correspondant en vert (à droite).*

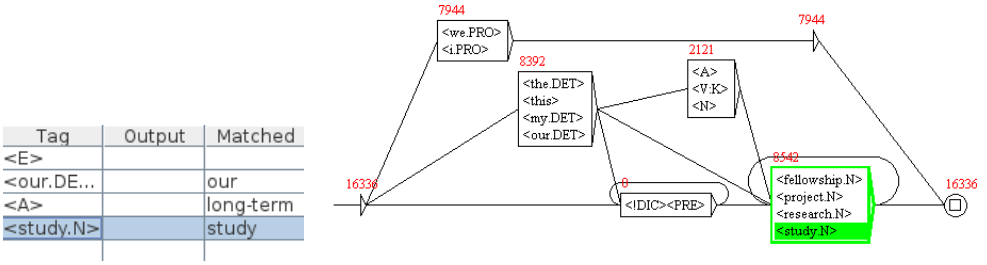


FIG. 14 – *Mot courant, « study », de la séquence parcourue (à gauche) et nœud (8542) du motif validé correspondant en vert (à droite).*

Tag	Output	Matched
<E>		
<our.DE...		our
<A>		long-term
<study.N>		study

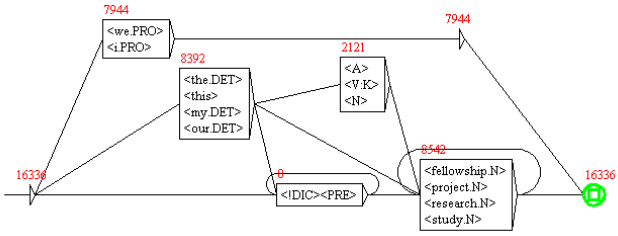


FIG. 15 – *Mot courant, fin de l'expression, de la séquence parcourue (à gauche) et nœud (16336) du motif validé correspondant en vert (à droite).*