Les expressions polylexicales transdisciplinaires dans les articles de recherche en sciences humaines : retour d'expérience (Chapitre 4)

Agnès Tutin

Chapitre 4

Les expressions polylexicales transdisciplinaires dans les articles de recherche en sciences humaines : retour d'expérience[1]

4.1. Introduction

Dans le cadre du projet ANR TermITH (2013-2016), présenté dans l'introduction de l'ouvrage (chapitres 1 et 5), le laboratoire LIDILEM a été chargé de constituer les ressources lexicales du lexique transdisciplinaire des sciences humaines. Une telle entreprise devait nécessairement prendre en compte, à côté du lexique des « mots simples », les expressions polylexicales (désormais EPL) sous toutes leurs formes, pour plusieurs raisons. Comme signalé dans le chapitre 1, de nombreux travaux ont montré que la phraséologie, définie comme l'ensemble des expressions préfabriquées, est extrêmement présente dans ce genre de texte très spécifique (voir par exemple [GLE 00, PEC 04, GRAN 06, DRO 07a, SIM 10, MAR 12]). Par ailleurs, la perspective phraséologique, en particulier à travers les collocations, donne à voir la dimension syntagmatique du lexique, indispensable à l'analyse sémantique des unités lexicales. En outre, la phraséologie étendue, à travers les routines, ces formules phrastiques qui renvoient à des fonctions discursives et rhétoriques spécifiques (par exemple [SIT 16, KRA 17]), constitue une voie d'entrée privilégiée pour l'étude de l'écriture scientifique et de l'épistémologie des disciplines. Enfin, s'il existe de nombreuses ressources phraséologiques des écrits scientifiques et académiques pour l'anglais (par exemple [SIM 10, MAR 12]), le français à l'heure actuelle ne dispose que de quelques ressources lacunaires dans ce domaine comme celle de Pecman [PEC 04].

Dans ce chapitre, les EPL qui nous intéressent ne concerneront pas la terminologie, mais uniquement le lexique transversal et « transdisciplinaire », qui décrit la démarche scientifique, la structuration textuelle et le raisonnement, c'est-à-dire des expressions comme *cas de figure, faire une hypothèse, comme on l'a vu/observé*, que l'on rencontre dans toutes sortes d'écrits scientifiques de sciences humaines (voir chapitre 1). Cette étude se veut avant tout un retour d'expérience sur un projet lexicographique visant des applications concrètes, comme le traitement automatique des langues, en particulier dans le cadre de l'indexation automatique (voir chapitre 5), ou la didactique des langues (voir chapitre 9). Nous présenterons ainsi les traitements et les choix effectués pour analyser un large ensemble d'expressions, ainsi que certaines expérimentations qui nous paraissent intéressantes. Ce travail a aussi été l'occasion pour nous de vérifier la faisabilité de certains traitements linguistiques sur un large ensemble phraséologique comprenant plusieurs types d'expressions.

Le chapitre s'organisera comme suit. Nous exposerons tout d'abord une typologie des expressions à l'œuvre dans le discours scientifique, puis nous nous focaliserons sur trois principaux types d'expressions polylexicales, dont nous observerons à la fois les techniques d'extraction et les modélisations. Nous nous intéresserons ainsi successivement : a) aux locutions, expressions polylexicales non compositionnelles (par exemple : *point de vue, tenir compte*), b) aux collocations binaires construites autour des mots simples du LST (par exemple : *jouer un rôle, hypothèse de départ, analyse critique*), et c) aux routines sémantico-rhétoriques, séquences stéréotypées associées à des fonctions rhétoriques et discursives spécifiques (par exemple : *comme on l'a vu/observé/constaté*…), que nous aborderons ici à travers l'exemple des routines à fonction métatextuelle.

[1]. Chapitre rédigé par Agnès TUTIN.

4.2. Typologie des expressions polylexicales adoptée dans le lexique TermITH

Le projet lexicographique que nous avons mené a principalement consisté à extraire et à décrire les différents types d'EPL. Il suppose donc d'établir entre les types d'EPL des frontières bien tracées, de façon à proposer des modélisations adaptées à chaque type d'élément. Ainsi, une EPL au fonctionnement compositionnel ne recevra pas le même traitement qu'une EPL plus figée, proche sur le plan sémantique d'un mot simple ; pour la routine sémantico-rhétorique, le fonctionnement discursif et rhétorique devra être précisé. Les choix effectués s'accommodent ainsi parfois mal de la complexité du continuum phraséologique – une expression taxée de collocation peut apparaitre plus « figée » qu'une autre de la même classe –, mais ces difficultés sont inhérentes à la perspective lexicographique.

La typologie des EPL adoptée dans le cadre du projet TermITH est plus rudimentaire que d'autres propositions (par exemple [GRA 08b, MEL 13, TUT 14a]). Elle dissocie, dans une distinction maintenant classique en phraséologie, les expressions lexicalisées – ou locutions – des collocations. Comme cela a déjà été rappelé (voir par exemple [LEG 13]), cette opposition était déjà présente chez Bally dans son *Traité de stylistique* (1909) [BAL 09], qui distinguait les « séries phraséologiques » (nos collocations) des « blocs indécomposables » (les « locutions »). Cette distinction n'est pas seulement théorique : elle a aussi des incidences concrètes puisque les deux types d'expression ne reçoivent pas le même traitement lexicographique. Les locutions, par leur non-compositionnalité, seront traitées comme des mots simples et recevront ainsi un traitement sémantique associé à la totalité de l'expression. Dans cette optique, *point de vue* sera traité sur le plan sémantique de la même façon que le nom simple *opinion*. En revanche, les collocations (par exemple : *analyse critique*) ne seront pas considérées comme des « unités lexicales », mais bien comme des associations lexico-sémantiques compositionnelles, dont chaque élément pourra recevoir une étiquette sémantique. Enfin, le troisième type d'EPL qui nous intéresse, les routines sémantico-rhétoriques (par exemple, *il est important/crucial de noter/observer…*), correspond à des objets linguistiques plus émergents, au croisement du lexique et du discours, dont la modélisation reste à affiner et dont nous proposons ici quelques pistes de traitement.

4.3. *Point de vue, mettre au jour, bel et bien* et autres locutions transdisciplinaires

Comme précisé précédemment, ont été considérées comme locutions les EPL présentant une faible compositionnalité sémantique, dans la mesure où le sens de l'expression apparaissait difficilement déductible des parties qui la composent [MAR 97]. C'est ainsi le cas de séquences comme *point de vue*, *mettre au jour* ou *bel et bien*, pour lesquelles le sens de l'ensemble apparait peu prédictible à partir des composants qui n'ont, d'ailleurs, pas d'autonomie référentielle (*jour* dans *mettre au jour* ne renvoie ici à aucun jour). Contrairement à d'autres auteurs (par exemple Gaston Gross [GRO 96]), nous n'avons pas dissocié dans notre typologie les locutions des mots composés, pour deux raisons : d'une part, il nous apparait difficile de différencier dans la pratique ces deux types ; d'autre part, une telle distinction n'a pas d'incidence sur le traitement lexicographique dans la mesure où locutions et mots composés seront traités de façon comparable, comme des blocs lexicaux. Par ailleurs, précisons que le critère de non-compositionnalité a été appliqué de façon assez lâche, dans la mesure où nous avons intégré dans la classe des locutions des expressions qui mettaient en jeu une part de compositionnalité, mais renvoyaient à un référent et à une dénomination précis. Par exemple, une EPL comme *étude de cas*, bien que sémantiquement motivée, a été intégrée dans la classe des locutions dans la mesure où ce terme désigne un type d'étude bien spécifique dans la recherche scientifique.

4.3.1. *L'extraction des locutions*

La première étape dans l'identification des EPL consiste à les repérer dans les textes, en extrayant les unités les plus représentatives. Dans le cadre de cette étude, c'est le corpus Transdisciplinaire TermITH, décrit aux chapitres 1 et 2, qui a été utilisé. Les techniques d'extraction des locutions à partir des corpus dépendent bien entendu des parties du discours concernées, dans la mesure où les locutions adverbiales ou nominales (par exemple : *en revanche, point de vue*) comportent généralement des éléments contigus, à la variabilité limitée, contrairement aux locutions verbales (par exemple. : *prendre en compte, mettre en évidence*), qui intègrent des

composants variables, non contigus et susceptibles d'alternances syntaxiques[2]. L'extraction des locutions à partir du corpus TermITH analysé syntaxiquement a reposé sur plusieurs techniques parallèles, utilisant toutefois les mêmes seuils de fréquence et de dispersion[3] (au moins 15 occurrences dans 3 disciplines sur 10). La première technique a simplement consisté à extraire des EPL déjà identifiées dans le lexique au même titre que des mots simples[4]. Cela a souvent été le cas des locutions adverbiales comme *en revanche* ou *sans cesse*. La deuxième méthode a consisté à utiliser des n-grammes ou segments répétés, c'est-à-dire des suites de mots contigües[5]. Cela nous a permis d'extraire des expressions comme *analyse du discours* ou *mettre en lumière*. Enfin, nous avons complété la liste des locutions extraites avec les techniques d'extraction des collocations, décrites à la section suivante. Les EPL extraites ont bien entendu été filtrées manuellement et l'examen des occurrences dans les textes a souvent conduit à rejeter des expressions qui ne répondaient pas aux critères fixés pour le lexique transdisciplinaire (voir chapitres 1 et 2).

4.3.2. *Le traitement des locutions*

Le traitement linguistique des locutions est analogue à celui des mots simples (voir chapitre 2). Ces EPL reçoivent également une glose, une étiquette de classe sémantique et de sous-classe sémantique. Un paramètre spécifique pour les variantes graphiques est ajouté pour cette classe d'éléments, certaines EPL ayant une orthographe instable (par exemple, *compte rendu* vs *compte-rendu*). Nous présentons ainsi dans le tableau 4.1 quelques exemples de locutions et leur traitement. Par exemple, l'adjectif *de taille* (comme dans *une différence de taille demeure*) recevra une glose extraite d'une ressource lexicographique (ici le *Dictionnaire Électronique des Mots* de Dubois et Dubois-Charlier [DUB 10]), accompagnée de deux étiquettes, pour la classe et la sous-classe sémantique (voir chapitre 2 pour une explication de ces classes).

Caté-gorie	Lemme	Variante graphi-que	Acception TermITH	Glose	Source	Classe sémantique	Sous-classe sémantique
A	de taille		de taille	Considérable, très important.	DEM	importance	importance_positif
Adv.	dès lors		dès lors_1	À partir de ce moment.	DEM	temporalité	chronologie
Nom	compte-rendu	compte rendu	compte-rendu	Rapport, exposé, ou relation de certains faits …	Wiktion-naire	communication_support	document
V	faire face		faire face	Affronter, s'opposer.	DEM	relation	#opposition

Tableau 4.1. *Exemples de locutions transdisciplinaires*

[2]. Par exemple, l'insertion d'éléments : *Nous avons pris ce problème en compte*..

[3]. Précisons ici que nous n'avons pas employé de seuil de spécificité pour l'extraction des locutions, car d'une manière générale, l'extraction des expressions polylexicales n'est souvent pas fiable, de nombreuses expressions polylexicales pouvant aussi avoir une lecture littérale.

[4]. L'analyseur utilisé repère déjà dans son lexique certaines locutions au même titre que des mots simples. Par exemple, la suite *tout à fait* est analysée comme *entièrement* comme un seul mot par le système. En revanche, *analyse du discours* est considéré comme une suite de 3 mots.

[5]. L'extraction des n-grammes a été effectuée avec un script développé par Olivier Kraif, que nous remercions ici.

À l'issue des traitements effectués, le lexique TermITH comporte 18 locutions adjectivales, 61 locutions verbales, 49 locutions nominales et 274 locutions adverbiales, dont certaines correspondent à plusieurs acceptions. Au total, les locutions transdisciplinaires sont finalement assez peu fréquentes, en dehors de la catégorie adverbiale qui est majoritairement composée de locutions.

4.4. *Liens étroits, jouer un rôle, hypothèse de départ* et autres collocations transdisciplinaires

Les collocations permettent d'observer les associations syntagmatiques privilégiées des mots du lexique transdisciplinaire. Leur extraction recourt généralement à la syntaxe et à des mesures statistiques. Leur traitement linguistique est toutefois plus complexe que les locutions, car il ne s'agit pas véritablement d'unités lexicales, mais pour la plupart d'entre elles d'associations lexicales.

4.4.1. *Extraction des collocations*

L'extraction automatique ou semi-automatique des collocations constitue un domaine maintenant bien connu du Traitement Automatique des Langues et de la lexicographie électronique (pour une synthèse assez complète, voir [EVE 07, DRO 07a]). La méthode choisie dans ce projet exploite à la fois la structure syntaxique reliant les éléments de la collocation et des mesures d'association statistique. Elle prend en compte par exemple le lien syntaxique entre le verbe *faire* et son objet direct le nom *hypothèse*, et privilégie les associations statistiques significatives. Cette méthode nous parait préférable aux techniques plus rustiques exploitant une fenêtre de mots et un étiquetage morphosyntaxique[6], qui génèrent davantage de bruit (voir aussi [SER 11]). On sait en effet que les associations syntaxiques mises en jeu dans les collocations peuvent parfois apparaitre à des distances éloignées dans le texte, qui ne seraient pas prises en compte dans une fenêtre de quelques mots, comme le montre l'exemple suivant qui relie *jouer* à *rôle* à une distance de 7 mots[7].

> (1) L'invention de l'écriture se serait-elle faite en deux étapes et le codage des nombres aurait-il joué, selon l'intuition d'André Leroi-Gourhan, un rôle d'initiateur ? [Article, Anthropologie]

Les collocations ont été extraites à partir de configurations syntaxiques, en utilisant l'outil Lexicoscope développé par Olivier Kraif et Sascha Diwersy [KRA 14, KRA 16]. Les paramètres suivants ont été utilisés pour extraire automatiquement les collocations [TUT 15], qui ont ensuite été filtrées manuellement :

– La base (ou élément stable) de la collocation appartient nécessairement à la liste des mots simples du LST [HAT 16b], par exemple *hypothèse* ou *analyser*. Dans de nombreux cas, les deux termes de la collocation relèvent du LST.

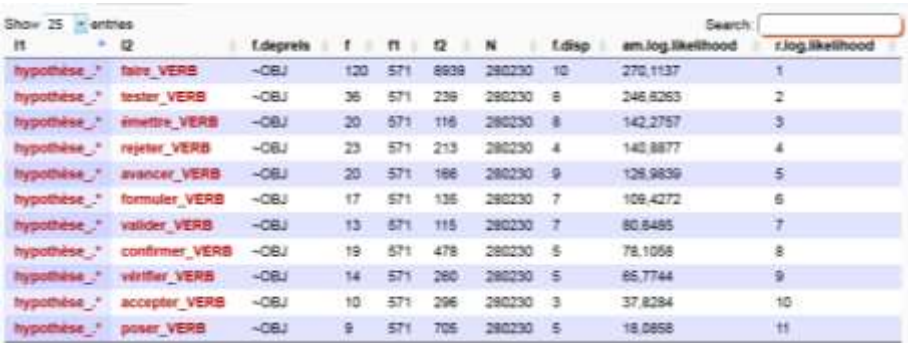
– Plusieurs seuils statistiques sont utilisés : une fréquence totale de la cooccurrence dépassant 7 occurrences, un rapport de vraisemblance (*log-likelihood ratio*) supérieur ou égal à 10.7 et une dispersion dans au moins 3 disciplines sur 10, de façon à garantir le caractère transdisciplinaire de l'association.

La figure 4.1 indique le résultat de l'extraction des verbes apparaissant en cooccurrence avec le nom *hypothèse* comme objet direct, en utilisant les paramètres d'extraction mentionnés.

---

[6]. Par exemple, le verbe *faire* associé au mot *hypothèse* dans une fenêtre de 6 mots.

[7]. Les relations syntaxiques permettent également de repérer des inversions syntaxiques, par exemple *l'hypothèse qu'il a faite.*



Show 25 entries Search:

f1	f2	f.deprels	f	f1	f2	N	f.disp	am.log.likelihood	r.log.likelihood
hypothèse_*	faire_VERB	~OBJ	120	571	8939	280230	10	270,1137	1
hypothèse_*	tester_VERB	~OBJ	36	571	239	280230	8	246,8263	2
hypothèse_*	émettre_VERB	~OBJ	20	571	116	280230	8	142,2757	3
hypothèse_*	rejeter_VERB	~OBJ	23	571	213	280230	4	140,8877	4
hypothèse_*	avancer_VERB	~OBJ	20	571	186	280230	9	128,9839	5
hypothèse_*	formuler_VERB	~OBJ	17	571	135	280230	7	109,4272	6
hypothèse_*	valider_VERB	~OBJ	13	571	115	280230	7	80,6485	7
hypothèse_*	confirmer_VERB	~OBJ	19	571	478	280230	5	78,1058	8
hypothèse_*	vérifier_VERB	~OBJ	14	571	260	280230	5	65,7744	9
hypothèse_*	accepter_VERB	~OBJ	10	571	296	280230	3	37,8284	10
hypothèse_*	poser_VERB	~OBJ	9	571	705	280230	5	18,0858	11

Figure 4.1. *Extraction, à l'aide du Lexicoscope, des cooccurrences lexico-syntaxiques des verbes ayant* hypothèse *comme objet direct*[8]

Nous avons pris en considération les constructions syntaxiques les plus productives de ces expressions lexicales, en élargissant les structures décrites par Hausmann [HAU 89] et en nous inspirant des configurations mises en œuvre dans les Fonctions Lexicales de la *Lexicologie Explicative et Combinatoire* [MEL 95, TUT 10a].

Ces schémas de collocations, symbolisés par des relations de dépendance dans les flèches (puisque l'analyseur syntaxique utilisé[9] adopte ce modèle), sont les suivants :

- N – PREP –> N : hypothèse de travail, formulation d'une hypothèse ;
- N – MOD –> A : hypothèse valide, mutation profonde, immense majorité ;
- V – SUJ –> N : (les) hypothèses confirment, (le) changement intervient ;
- V – OBJ –> N : faire une hypothèse, jouer un rôle ;
- V – PREP –> N : s'appuyer sur une hypothèse, aboutir à un résultat ;
- V –MOD –> ADV : assumer pleinement, exclure totalement, modifier profondément ;
- A – MOD –> ADV : totalement absent, significativement différent ;
- N – PREPOBJ –> PREP : grâce à l'analyse, selon l'approche.

Une fois l'extraction réalisée à l'aide des critères présentés ci-dessus, un filtrage manuel a été effectué pour écarter les associations terminologiques propres à un champ disciplinaire (par exemple, *capacité de production* ou *règle pragmatique*), les expressions qui relèvent plutôt des objets des sciences humaines (comme *cadre de vie* ou *secteur d'activité*, voir chapitre 2) ou, dans certains cas, des expressions qui ont bien un caractère transdisciplinaire, mais s'apparentent davantage à des locutions comme *point de vue*, et qui ont été traitées dans la section 4.3. Les expressions retenues ne sont pas nécessairement très idiomatiques ; la plupart des collocations retenues peuvent être considérées comme des collocations « régulières » [TUT 02], c'est-à-dire répondant à des schémas sémantiques compositionnels, même si elles n'apparaissent pas forcément prédictibles pour le locuteur non natif, surtout s'il est peu familiarisé avec les écrits scientifiques. Rappelons que l'objectif de notre lexique est avant tout de répondre à des besoins applicatifs, aussi bien pour le traitement automatique des langues que pour la didactique du français scientifique.

Par ailleurs, sur le plan sémantique, la plupart des collocations relevées répondent à une structure de type prédicat-argument [TUT 13], le collocatif ayant une fonction de prédicat, la base une fonction d'argument. Dans *absence totale* par exemple, l'intensificateur *total* est le prédicat qui porte sur la base *absence*. Toutefois, certaines associations lexicales, que nous avons décidé d'inclure dans la classe des collocations, ont davantage

---

[8]. Le tri des expressions est effectué par ordre décroissant de la mesure du rapport de vraisemblance. « F.deprels » indique la relation syntaxique de dépendance, ici *objet direct*. « f » : fréquence de la co-occurrence. « f1 » : fréquence du mot 1, ici *hypothèse*. « f2 » : fréquence du mot 2, ici *faire*. « f.disp » : dispersion dans les 10 disciplines. « Am.log.likelihood » : mesure d'association, ici *rapport de vraisemblance*. « R.log.likelihood » : rang dans le classement par ordre décroissant du rapport de vraisemblance.

[9]. Ici, Xerox Incremental Parser.

une fonction « classifiante » que « qualifiante » (voir [DIA 17]). Nous intégrons ainsi des expressions comme *analyse statistique* ou *recherche fondamentale* dans la catégorie des collocations.

4.4.2. *Traitement linguistique des collocations*

Le traitement proposé pour les collocations est assez simple : il prévoit un accès à la collocation à la fois par la base ou le collocatif, contrairement aux dictionnaires « papier » de collocations, qui privilégient la base. La microstructure de l'article traitant la collocation intègre en effet un ensemble d'informations qui doivent être manipulées par une base de données.

Nous détaillons la microstructure de l'article de la collocation, et l'illustrons à l'aide de la collocation *figure centrale*, comme dans l'exemple suivant :

(2) Parce qu'il est une figure centrale de la dissidence et parce qu'il a été accusé d'espionnage pour le compte des États-Unis, Chtcharanski aurait dû bénéficier d'une mobilisation sans précédent de Washington. [Article, Histoire]

– Les éléments constitutifs de la collocation

Comme nos collocations mettent en jeu des structures binaires, la première information concerne les deux éléments de la collocation, en particulier la base et le collocatif, dont nous indiquons la forme lemmatisée, mais aussi l'acception retenue dans la base TermITH des mots simples (voir chapitre 2)[10]. Dans notre exemple, la collocation est constituée de deux lemmes, *figure* relié à l'acception figure_1, *central* relié à central_2. ('figure' en tant que symbole et non en tant que schéma, 'central' au sens d''important' et non au sens spatial). La collocation *figure centrale* sera reliée à ces deux mots simples dans la base de données.

– Forme de base de la collocation

La forme de base indique la forme de surface la plus productive pour la collocation. Ainsi, dans les associations de type N-MOD->A, l'adjectif peut être préférentiellement postposé (*figure centrale*) ou antéposé (*vif débat*). Dans les collocations de type N-PREP->N, un déterminant peut apparaitre ou non devant le nom régi, phénomène étroitement lié au fonctionnement sémantique de l'association (Cf. [GRO 16] sur ce sujet). Les questions du nombre des noms, ainsi que du type de déterminant, sont également à prendre en compte. Pour l'exemple qui nous intéresse, la forme de base *figure centrale* sera proposée.

– Structure syntaxique de la collocation

La microstructure intègre également le type de structure syntaxique, modélisé à l'aide de relations syntaxiques de dépendance. Dans le cas présent, il s'agit de la relation N–MOD–>A, qui relie un nom à un modifieur adjectival.

– Fonction Lexicale (FL) éventuelle

Comme de nombreuses collocations correspondent à des schémas sémantiques productifs, nous avons indiqué les fonctions lexicales standard de la *Lexicologie Explicative et Combinatoire* [MEL 95], lorsque cela était pertinent, principalement avec les fonctions lexicales suivantes :

- Magn pour l'intensification : absence totale, nombre important, totalement absent, assumer pleinement ;
- $Oper_n$ pour les constructions à verbe support : *faire une comparaison, jouer un rôle* ;
- CausFunc pour les causatifs : donner la priorité, donner la possibilité.

L'attribution de ces FL à un grand nombre de collocations de notre lexique montre la productivité de cette modélisation pour un lexique à fonction scientifique.

[10]. Certaines structures syntaxiques, en particulier les structures N de N, peuvent toutefois être parfois difficiles à analyser dans le cadre de cette dichotomie base-collocatif (voir [TUT 13]).

Le tableau 4.2 résume le traitement effectué pour *figure centrale*.

Lemme de la base	figure
Lemme du collocatif	central
Acception de la base	figure_1
Acception du collocatif	central_2
Forme de base	figure centrale
Structure de la collocation	N-MOD->A
Fonction lexicale	_

Tableau 4.2. *Traitement de la collocation* figure centrale

La base TermITH des collocations contient 2 130 EPL réparties dans les schémas collocationnels indiqués plus haut. D'autres traitements seraient intéressants à effectuer de façon à mieux rendre compte du comportement des collocations, dont on sait qu'elles sont de manière générale bien plus variables que les locutions. Pour les collocations verbales, il serait pertinent de prendre en compte de façon systématique les sous-catégorisations et les alternances (voir [TUT 15, DAN 16]), comme on l'observe dans le tableau 4.3, qui présente les variations de sous-catégorisation et d'alternances de la collocation *faire une hypothèse*. Par manque de temps dans le cadre du projet, cette étude n'a toutefois été effectuée que sur un sous-ensemble d'expressions.

Collocation	Paramètres syntaxiques	Faire-hypothèse	Exemples
Déterminants		l' (91 %), des (8 %), une (0,5 %)	
Alternances syntaxiques	objet direct	92 %	Nous avons fait l'hypothèse.
	passif	2 %	Ainsi, aucune hypothèse n'est faite sur le caractère…
	passif réduit	2 %	Les conséquences des hypothèses générales faites sur l'illocutoire.
Sous-catégorisation externe	Que-P	(avec l') : 58 %	On peut donc faire également l'hypothèse que c'est au cours de la Belle Époque que s'opère la transition.
	Prep-de	(avec l') : 12 %	On peut faire l'hypothèse d'une forte homogénéité des sujets…

Tableau 4.3. *Variabilité syntaxique de la collocation* faire-hypothèse

4.5. Routines sémantico-rhétoriques : l'exemple des routines métatextuelles

Nous présenterons enfin un dernier type d'expression, les routines sémantico-rhétoriques, dont nous avons exploré le fonctionnement dans le cadre du projet TermITH, mais qui n'ont pas encore fait l'objet d'une description complète. Plusieurs travaux récents se sont intéressés à ce type d'expressions sous le terme de motifs (voir [LON 13]), ou routines [NEE 14] (pour une synthèse, voir le numéro 53 de *LIDIL* [SIT 16] et le récent numéro de la revue *CORPUS* [BEN 17]). Dans ce travail, nous définissons les routines comme répondant aux caractéristiques suivantes (voir [TUT 16, KRA 17], pour une définition plus détaillée) : a) ce sont des motifs récurrents, b) ayant une fonction discursive et rhétorique propre à un genre textuel, c) mettant en jeu une configuration lexico-syntaxique spécifique avec des paradigmes lexicaux, et d) correspondant à un énoncé actualisé dans le texte, c'est-à-dire renvoyant à des référents spécifiques comme l'objet d'étude, l'énonciateur du texte ou son destinataire[11].

À titre d'exemple, nous présenterons ici les routines à fonction métatextuelle, c'est-à-dire les routines qui renvoient à la navigation textuelle et à la structuration du texte, et qui sont des éléments discursifs qui apparaissent centraux dans les écrits scientifiques et argumentatifs. Nous exposerons tout d'abord le processus d'extraction, qui exploite la méthode des arbres lexico-syntaxiques récurrents associée ici aux classes sémantiques du LST TermITH des mots simples (décrit aux chapitres 1 et 2), puis les résultats de ces extractions.

4.5.1. *Extraction des routines sémantico-rhétoriques*

Pour extraire les routines, nous exploitons une méthode combinant des techniques statistiques à des associations syntaxiques, un peu à la façon de n-grammes, mais en exploitant des relations syntaxiques de dépendance [TUT 16, KRA 17]. Cette méthode, qui extrait des arbres lexico-syntaxiques récurrents (ALR), a été comparée à l'approche plus surfacique des n-grammes [KRA 17] et a montré sa pertinence pour l'extraction des routines. Nous présentons ici une expérimentation à l'aide des classes sémantiques du LST TermITH de façon à extraire non de simples suites de configurations lexico-syntaxiques, mais des structures syntaxiques et sémantiques, plus abstraites. L'extraction des ALR fonctionne de façon itérative, en utilisant des associations binaires qui servent ensuite de pivots à de nouvelles extractions (pour une présentation plus précise, voir [KRA 17]). Si nous lançons ainsi l'extraction des ALR sur les verbes de formulation, nous extrayons ainsi un arbre syntaxique comme celui de la figure 4.2, qui repère les cooccurrences syntaxiques de mots et de classes sémantiques.

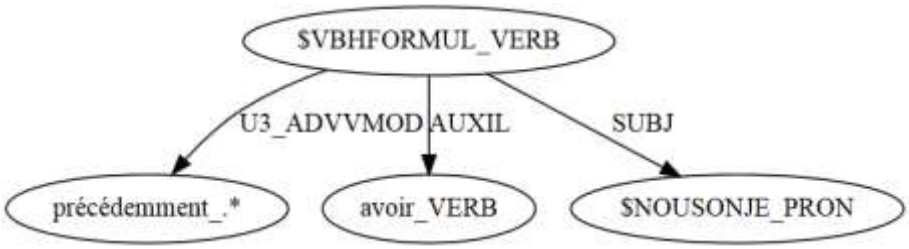


Figure 4.2. *Arbre lexico-syntaxique exploitant des classes sémantiques du lexique TermITH*

La classe sémantique des verbes de formulation $VBHFORMUL (*dire, indiquer, mentionner,…*) se combine ainsi avec des pronoms sujets qui renvoient à l'auteur $NOUSONJE (*nous, on, je*) et avec l'adverbe *précédemment*. Cet ALR correspond ainsi aux exemples suivants dans le texte :

(3) En effet, comme nous l'avons évoqué précédemment, une phase préalable à l'étude proprement dite avait indiqué que la corrélation […]. [Article, Psychologie]

[11]. Les collocations constituent un matériel lexical abstrait. Les routines, par opposition, mettent en jeu des référents identifiés dans le discours. Par exemple, dans la routine *comme on l'a vu/observé*, le *on* renvoie à l'auteur et aux destinataires.

(4) On a précédemment indiqué que la présence d'une mémoire longue dans les séries de rentabilités impliquait une certaine prévisibilité des rentabilités futures à partir des rentabilités passées. [Article, Économie]

L'analyseur employé utilisant les dépendances, on notera ici que l'adverbe peut indifféremment succéder au verbe ou le précéder. L'ALR extrait correspond bien à une routine métatextuelle qui rappelle des faits antérieurs mentionnés dans le même document ou d'autres travaux du même auteur.

Pour le repérage des routines à fonction métatextuelle, plusieurs classes sémantiques du LST TermITH (voir chapitres 1 et 2) ont été utilisées. Nous avons tout d'abord exploité le lexique associé à la structuration textuelle, qu'il s'agisse de marqueurs de navigation textuelle[12] (*infra, ci-dessous, suivant, précédent,…*) ou de marqueurs d'intégration linéaire [JAC 03] (*dans un premier temps, d'autre part, pour finir*,…). Nous avons également exploité les noms renvoyant aux supports de l'écrit scientifique, qu'il s'agisse des documents (*article, rapport, document, texte…*), des parties des documents (*section, chapitre, annexe…*) ou des supports graphiques (*tableau, figure, schéma…*). Enfin, une série lexicale renvoie au processus de formulation scientifique, dans sa dimension textuelle et linguistique (*mentionner, préciser, dire, mention, illustrer, détailler …*). Les classes sémantiques comportent également quelques mots fréquents, comme le verbe *dire*, qui n'appartiennent pas au LST-TermITH, car trop courants dans le corpus de référence, mais ces mots nous paraissent pertinents pour l'extraction des routines. Un seuil de fréquence à 10 pour chaque itération[13] et une dispersion de 3 disciplines sur 10 sont utilisés[14].

4.5.2. *Caractérisation des routines textuelles*

Les routines extraites à l'aide des classes sémantiques et de la méthode des ALR sont à peu près au nombre d'une trentaine, une fois le filtrage manuel effectué[15]. Pour les modéliser, il sera important d'indiquer la ou les fonctions rhétoriques et discursives, ainsi que la structure sémantique et syntaxique de la routine, qui intègre des paradigmes lexicaux dans la plupart des cas. Dans ce qui suit, nous présenterons principalement les principales routines rhétoriques mises en évidence à l'aide des classes sémantiques présentées plus haut, en nous focalisant sur les fonctions textuelles. Les routines à fonction métalinguistique (*si l'on peut dire, pour reprendre l'expression/la formulation* de X) n'ont pas été ici prises en considération.

4.5.2.1. *Les renvois intertextuels*

L'écrit scientifique est un genre textuel où le dialogisme est très présent, en premier lieu dans sa dimension interdiscursive. La connaissance scientifique exposée se tisse à travers un ensemble de discours dont les références sont souvent explicites (à travers la mention auteur-date, en particulier). Certaines routines ont explicitement pour fonction de renvoyer le lecteur à d'autres références :

Pour une [présentation/synthèse] ([détaillée/brève]), [cf./voir/se reporter] à [AUTEUR-DATE]

(5) Pour une présentation détaillée de ce concept, voir la note de synthèse parue dans la Revue française de pédagogie (Sarrazy, 1995). [Article, Sciences de l'éducation]

(6) Pour une présentation de la théorie des files d'attente, voir KLEINROCK [1975]. [Article, Économie]

On notera l'emploi quasi exclusif de ces formules dans les notes de bas de page, le verbe de renvoi étant généralement à l'infinitif injonctif. À côté de ces pointeurs injonctifs, on relève des renvois intertextuels intégrés au sein du texte principal, qui mettent davantage en avant l'expertise de l'auteur, souvent sur un point spécifique.

---

[12]. Les résultats pâtissent du fait qu'un large sous-ensemble d'expressions reste encore mal reconnu par l'analyseur qui utilise XIP. L'analyseur est toutefois un excellent système d'analyse syntaxique.
[13]. Le seuil pour le rapport de vraisemblance utilisé dans les associations syntaxiques est de 10,81.
[14]. Les seuils sont ici un peu plus bas que pour les n-grammes, du fait des itérations.
[15]. Nous ne conservons que les routines qui correspondent à des propositions. De nombreux éléments extraits par cette technique sont des collocations.

On peut citer/mentionner X

(7) Les travaux sur ce mot sont peu nombreux, on peut citer l'étude lexicographique de M. Herman (1992) dans les *Cahiers de lexicologie* en 1992, un article de L. Tracy dans *Langue française* (1997) et un autre de S. Meleuc paru dans *LINX* en 1999. [Article, Sciences du langage]

4.5.2.2. *Les renvois intratextuels*

L'écrit scientifique est un écrit long et complexe, qui nécessite des marques de guidage pour le lecteur. Des renvois intratextuels sont ainsi proposés au lecteur pour mieux repérer et parcourir le support textuel. Le renvoi peut être contextuel (*ci-dessus, supra*) ou absolu (*section 2, annexe B*). Le motif apparait généralement entre parenthèses et recourt au verbe *voir* injonctif ou *cf.*

[Cf./voir] [supra/infra/ci-dessus/ci-dessous…]

[Cf./voir] [section/chapitre/annexe…] N

(8) D'autre part, avec l'étiquette de « phrase prototypique », le problème semble déserter l'objectivité des faits pour une forme de subjectivité propre au jugement d'appartenance catégorielle (voir infra). [Article, Sciences du langage]

À côté de ce pointeur intratextuel, on observe des formulations qui visent aussi à mettre en avant la cohérence textuelle interne, tout en effectuant des renvois (plus ou moins explicites) à d'autres parties textuelles. Ces formulations apparaissent souvent dans des énoncés parenthétiques (voir [GRO 14a]).

[je/on/nous] <avoir> [évoqué/mentionné/vu] [précédemment/plus haut]

comme évoqué précédemment/plus haut

comme [je/nous/on] l' <avoir> [dit/évoqué/mentionné]

(9) […] comme on l'a déjà évoqué, Schaeffner part en mission avec une formation essentiellement musicale (on se souvient par exemple de l'importance du solfège rythmique dans son apprentissage) et se retrouve intégré dans un protocole d'enquête ethnographique. [Article, Anthropologie]

4.5.2.3. *La structuration du texte*

Un ensemble de routines vise aussi à faciliter la lecture du texte en présentant au lecteur la structure interne du document. Ces marques emploient fréquemment un marqueur d'intégration linéaire dans une structure énumérative [HOD 10] (*tout d'abord, en premier lieu, pour finir*), un verbe de description et la mention de l'auteur du texte.

[je/on/nous] [exposer/proposer/présenter/décrire/détailler…] [en premier lieu/premièrement/dans un premier temps/tout d'abord/pour finir…]

(10) La stratégie retenue pour présenter ces résultats est la suivante. Dans un premier temps**,** nous exposons une structure du marché de l'assurance chômage qui donne la possibilité aux ménages d'assurer complètement leurs choix de consommation contre le risque individuel de chômage. [Article, Économie]

À côté de ces éléments liés à la structuration, on repère des routines qui indiquent le « topique » principal de l'article (voir aussi [JAC 13]).

[objet/but] de cet article est de X

[on/nous/je] <proposer> dans cet article de X

cet article propose X

(11) Nous proposons dans cet article de présenter […] les conditions – notamment politiques et institutionnelles – de la diffusion de cette innovation. [Article, Sciences de l'information et de la communication]

4.5.2.4. *Les marques de topicalisation*

Enfin, un dernier type de routines particulièrement présent vise à mettre en relief des informations à destination du lecteur.

Il est [important/intéressant] de [signaler/souligner/noter/préciser] que

Il faut [signaler/souligner/noter/préciser] que

(12) À cet égard, il est important de noter que les observations concernant le vent ne répondaient pas à un désir de connaissance scientifique, mais à des considérations pratiques. [Article, Histoire]

Les routines présentées ici n'intègrent bien entendu pas la totalité des éléments associés aux fonctions métatextuelles, dont certaines peuvent échapper aux seuils statistiques proposés ou être formulées à l'aide de synonymes peu fréquents, exclus de nos classes sémantiques. La méthode des classes sémantiques associée à la technique des ALR parait toutefois plus prometteuse que les méthodes basées sur les associations strictement lexicales, en capturant les alternances lexicales et en proposant des schémas de routines plus abstraits. La modélisation des routines, qui n'est ici qu'esquissée, reste également à poursuivre. On a pu voir à travers les exemples présentés plus haut que le fonctionnement discursif et énonciatif des routines (position dans la phrase, distribution dans les parties textuelles, interaction avec le lecteur) était un point particulièrement important, qu'il faudra prendre en compte dans les descriptions.

4.6. Conclusion

Le retour d'expérience présenté dans ce chapitre illustre bien la diversité des EPL dans les écrits scientifiques. La ressource constituée dans le cadre du projet TermITH intègre déjà un ensemble d'EPL, accompagnées de traits sémantiques et syntaxiques. Les locutions figées, traitées sur le plan linguistique de façon analogue aux mots simples – ce sont des blocs indécomposables –, ne posent guère de difficulté, en dehors de leur repérage dans les textes qui reste délicat. Les collocations, qui ne sont pas pour la plupart de véritables unités lexicales, ont été extraites à l'aide d'une méthode classique qui exploite à la fois la structure syntaxique et des critères statistiques. Sur le plan lexicographique, chaque élément de la collocation est traité de façon distincte. Enfin, les routines sémantico-rhétoriques, qui sont des notions linguistiques émergentes, n'ont pas encore donné lieu à un inventaire systématique. La méthode d'exploration, exploitant les arbres lexico-syntaxiques récurrents et les classes sémantiques du lexique TermITH des mots simples, s'est révélée très prometteuse. L'application aux routines métatextuelles a montré l'intérêt de ce mode d'exploration, dont la modélisation doit encore être affinée, en particulier pour prendre en compte la distribution dans les phrases et dans les textes.

Ces ressources peuvent être exploitées avec profit dans le cadre d'applications de TAL. Pour l'indexation automatique (voir chapitre 5), les associations lexicales peuvent être davantage prises en compte, en excluant des candidats termes les associations transdisciplinaires. Par exemple, l'acception linguistique du mot *objet* sera écartée dans des collocations transdisciplinaires comme *faire l'objet* ou *l'objet du débat*. Les routines repérées peuvent également être utilisées lorsqu'elles permettent d'introduire des termes spécifiques, par exemple des routines définitoires *Nous appelons X/désignons par X…* [REB 01, JAC 11]. Les applications didactiques comme celles qui sont présentées dans ce volume (chapitres 7, 8 et 9) sont évidemment nombreuses. Une perspective particulièrement intéressante est celle de l'aide à la lecture de textes académiques (voir [LUN 13]) par le repérage des motifs récurrents.