Introduction

Le lexique scientifique transdisciplinaire

Dans cet ouvrage, nous souhaitons, plus de dix ans après le numéro de la *Revue française de linguistique* (*RFLA*) intitulé « Autour du lexique et de la phraséologie des écrits scientifiques » [TUT 07a], revenir sur la question du lexique scientifique transdisciplinaire, ce lexique transversal des sciences, largement partagé par les disciplines. Depuis le numéro de la *RFLA*, ce domaine a connu des évolutions notables, avec la parution en anglais de ressources phraséologiques de référence, comme l'*Academic Formulas List* ou l'*Academic Collocation List* (section I.3), et de nouveaux ouvrages comme celui de Paquot [PAQ 10]. Dans le cadre du projet ANR TermITH[1] (voir section I.5 et chapitre 4), nous avons pu, grâce aux travaux des chercheurs de l'ATILF et du LIDILEM[2], en particulier à travers plusieurs thèses [TRA 14, HAT 16a, YAN 17], constituer des ressources nouvelles et approfondir notre réflexion sur ce thème. Il nous parait donc opportun de dresser, dans cet ouvrage collectif, un bilan de nos réflexions.

La communication scientifique est une activité centrale des enseignants-chercheurs, comme des chercheurs en devenir, qui sont amenés à produire des discours écrits et oraux diversifiés [LEF 06] : communications informelles entre pairs ou par messagerie, ou productions écrites variées (articles de recherche, communications écrites, comptes rendus de lecture, monographies). Dans cet ouvrage, nous aborderons essentiellement le lexique du discours scientifique à travers un genre assez codifié, celui de l'article de recherche, genre emblématique de l'activité scientifique du chercheur, qui, par sa densité et son caractère argumentatif, constitue une forme de modèle pour les apprentis

---

Introduction rédigée par Agnès TUTIN et Marie-Paule JACQUES.

1. TERMinologie et indexation de textes en sciences humaines

2. ATILF : Laboratoire d'analyse et traitement informatique de la langue française ; LIDILEM : Laboratoire de linguistique et didactique des langues étrangères et maternelles.

chercheurs. En outre, nous nous intéresserons essentiellement aux sciences humaines et sociales (SHS), que nous ne chercherons pas à définir, au-delà du fait qu'elles ont pour objet l'homme dans sa dimension sociale. Nous partons du constat que les SHS constituent une famille de disciplines dans les classifications institutionnelles comme celles du CNRS[3] (INSHS) ou des universités[4]. Si nous nous intéressons aux SHS dans leur ensemble, il ne s'agit pas, bien entendu, de prétendre qu'elles forment un groupe parfaitement homogène, les critères de scientificité et les fonctionnements épistémologiques pouvant être assez différents d'une discipline à l'autre. Ce qui caractérise l'écriture scientifique en SHS, d'ailleurs beaucoup moins étudiée qu'en sciences expérimentales [GRO 10a], c'est l'importance qu'elle occupe dans l'activité scientifique du chercheur, ce qui en fait un enjeu crucial pour les apprentis chercheurs.

Parmi les caractéristiques linguistiques des écrits de recherche en SHS, le lexique occupe une place centrale. Nous rappellerons dans cette introduction ce que nous appelons le « lexique scientifique transdisciplinaire », puis en esquisserons les frontières avec les lexiques limitrophes. Nous présenterons ensuite quelques ressources lexicales pour le français et l'anglais, puis nous aborderons rapidement les méthodes mises en place pour l'élaboration du lexique scientifique transdisciplinaire (LST) dans le cadre du projet TermITH. Les finalités de TermITH seront ensuite abordées et le chapitre se clôturera par une présentation des contributions de ce volume.

I.1. Qu'est-ce que le lexique scientifique transdisciplinaire ?

I.1.1. *Les propriétés lexicales : des caractéristiques linguistiques parmi d'autres*

Le discours scientifique a fait l'objet de nombreuses recherches en linguistique, en particulier en langue anglaise (par exemple [SWA 90, HYL 04, HAL 06]), mais aussi dans d'autres langues comme le français (voir par exemple [KOC 82, GRO 10a, RIN 10, TUT 14c]). Les travaux linguistiques dans ce champ sont diversifiés (pour une synthèse, voir [RIN 10]), mais la question du lexique, qui nous intéresse ici, reste assez peu étudiée en dehors de la terminologie. C'est ce lexique non terminologique qui nous occupera, en particulier dans le domaine des SHS, moins étudiées sous cet angle. Notre objectif premier est de contribuer à une réflexion et à une description plus approfondies du lexique propre au genre des articles de recherche en sciences humaines et sociales, afin

3. Centre national de la recherche scientifique.

4. Comme en témoignent par exemple le regroupement d'un ensemble de disciplines dans l'Institut national des sciences humaines du CNRS (on notera que la psychologie n'y est rattachée que secondairement) ou l'organisation scientifique de certaines universités comme l'université Grenoble Alpes avec un « Pôle Sciences humaines et sociales ».

de constituer un matériau permettant des études plus systématiques des fonctionnements énonciatifs et rhétoriques de ces textes, de proposer une meilleure compréhension de l'épistémologie des disciplines et de fournir des outils pour l'enseignement du discours scientifique et académique ainsi que pour le traitement automatique des langues.

Le discours scientifique, en particulier dans les articles de recherche, se caractérise par un ensemble de propriétés linguistiques, sur les plans syntaxique, textuel et lexical, étroitement liées à des fonctions rhétoriques spécifiques. Sur le plan syntaxique, plusieurs spécificités ont été soulignées [KOC 82, HAL 06] : la prédominance de la modalité déclarative, le suremploi du présent, la fréquence des constructions incises, des constructions impersonnelles et passives, et le recours aux nominalisations [HAL 06]. L'exploitation de corpus annotés syntaxiquement permet maintenant de quantifier ces caractéristiques, comme l'étude de Fifielska [FIF 15] qui a comparé l'emploi de plusieurs types de passifs et de constructions impersonnelles d'un ensemble d'articles de recherche en sciences humaines avec un corpus journalistique. Elle a ainsi, entre autres, relevé un net suremploi (a) des constructions impersonnelles adjectivales comme *il est* Adj *de* V (*il est important/crucial de noter/mentionner*), (b) des structures impersonnelles passives (*il est admis que, il a été montré que*), souvent associées à une forme d'effacement énonciatif de l'auteur fréquent dans ce genre (voir par exemple [FLØ 06]).

Les aspects énonciatifs des écrits scientifiques ont également été beaucoup abordés (voir [RIN 10]), en particulier à travers la figure de l'auteur [FLØ 06]. Dans le cadre du projet Scientext et dans sa continuité, un ensemble d'études ont été menées sur les marques de positionnement et de raisonnement à travers le lexique évaluatif, l'emploi des verbes et des marques dialogiques [GRO 09, GRO 10c, TUT 10, HAR 14, TUT 14c].

Sur le plan textuel, également, d'autres spécificités se font jour. La structure textuelle présente une organisation très codifiée (avec titres et intertitres) [JAC 14], voire la contrainte du canevas IMRAD (*Introduction, Methods, Results and Discussion*), toutefois assez peu courante en SHS. On observe également plusieurs fonctionnements assez caractéristiques sur le plan de la cohésion textuelle, avec l'emploi fréquent d'anaphores résomptives, exploitant notamment le lexique abstrait des *shell nouns*[5] [FLO 15]. Cependant, comme dans toute langue de spécialité (voir le manuel récent de Forner et Thörle [FOR 16] ou Lerat [LER 97]), c'est bien au niveau du lexique employé que les caractéristiques linguistiques sont les plus manifestes, à travers les unités lexicales propres aux disciplines, mais aussi à travers le lexique qui aborde le raisonnement, la narration et l'activité scientifiques. C'est ce lexique de genre transversal qui fera l'objet de cet ouvrage.

5. Généralement traduit par « noms coquilles ».

I.1.2. *Du lexique disciplinaire au lexique transdisciplinaire*

Le lexique disciplinaire renvoie aux notions scientifiques délimitées par les disciplines. Pour prendre un exemple que nous pensons familier aux lecteurs, la linguistique, nous considèrerons comme appartenant à son lexique disciplinaire les concepts définis dans les dictionnaires des sciences du langage, qui relèvent de la terminologie du domaine. Par exemple, les premières pages du *Dictionnaire de linguistique et de sciences du langage* [DUB 99] définissent les termes suivants : *abduction*, *aberrant*, *abessif*, *ablatif*, en indiquant le domaine de la linguistique concerné par la notion, respectivement « en phonétique », « dans les langues finno-ougriennes », « en syntaxe ». Ce lexique de spécialité comprend évidemment de nombreux termes complexes (*achoppement syllabique*, *segment acoustique*, *adjectif verbal*…), essentiellement nominaux, mais aussi plus rarement quelques éléments verbaux ou adjectivaux. Ce lexique de spécialité est aussi employé dans les thésaurus des spécialistes de l'indexation, comme dans *Thésaulangue* pour la linguistique (voir chapitre 4).

Le lexique disciplinaire est évidemment très présent dans les articles de recherche, qu'il s'agisse de néologismes de forme (termes « unidisciplinaires » au sens de Kocourek [KOC 82], comme *ablatif* ou *allophone*) ou de néologismes sémantiques, peut-être plus fréquents. Ces derniers empruntent des formes de la langue générale et leur attribuent une acception spécifique, par exemple les notions de *compétence* ou de *distribution* en linguistique. Le lexique disciplinaire est essentiellement étudié par les terminologues, parfois par les morphologues et les spécialistes du traitement automatique des langues.

À côté de ce lexique disciplinaire, on observe un ensemble d'unités lexicales qui ne désignent pas véritablement des éléments techniques, mais qui renvoient à la démarche et aux activités scientifiques, au raisonnement et à l'écriture scientifique (voir [PEC 04, DRO 07a, TUT 07a] pour d'autres définitions). Ces éléments sont « transdisciplinaires » en ce qu'ils traversent en quelque sorte les disciplines et correspondent donc au lexique stable du genre, par opposition au lexique disciplinaire propre à une sphère scientifique particulière. Nous retrouverons ainsi dans la plupart des disciplines – des sciences humaines aux sciences les plus « dures » – des lexèmes comme *hypothèse, analyser, résultats,* mais aussi des expressions comme *dans un premier temps, pour conclure,* qui sont étroitement liées à la démarche scientifique et à l'écriture scientifique. Un petit sondage effectué dans le corpus Transdisciplinaire TermITH (corpus diversifié de sciences humaines, voir section I.4) a permis d'extraire quelques collocations fréquentes, largement partagées. Les résultats (voir la figure I.1) indiquent que deux de ces collocations (*faire une hypothèse, proposer une analyse*), bien que de distribution inégale, apparaissent dans les sous-corpus de toutes les disciplines, ce qui montre que les processus scientifiques dénotés sont largement transversaux aux disciplines des sciences humaines. De façon intéressante, les deux autres collocations qui renvoient aux méthodes

sont en revanche moins transversales, ce qui révèle des pratiques méthodologiques propres à des sous-ensembles de disciplines : *mener une enquête* est davantage présent dans les sciences sociales (en particulier en anthropologie), alors qu'*analyse statistique* est, sans surprise, fortement présent dans les disciplines les plus expérimentales des sciences humaines et sociales (en particulier, la psychologie et l'économie). Le repérage du lexique et des expressions communs permet ainsi de caractériser la démarche scientifique et les modes d'écriture partagés et, par la suite, de comparer les pratiques au sein d'un ensemble plus large.

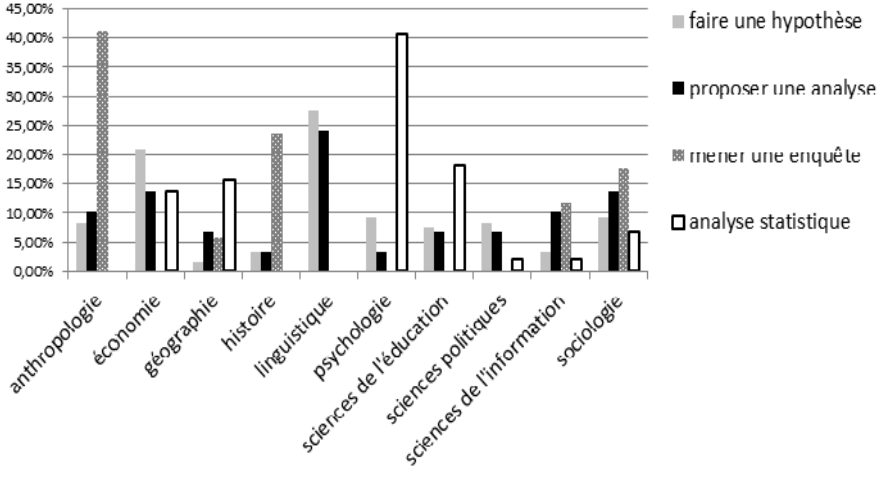

Figure I.1. *Fréquence et répartition de quelques collocations transdisciplinaires*

Bien entendu, il ne s'agit pas de prétendre que toutes les disciplines scientifiques utilisent le même lexique de genre, les démarches et les pratiques scientifiques variant largement au sein des familles disciplinaires. Hyland et Tse [HYL 07] ont ainsi bien montré, dans une étude de corpus, que le lexique transdisciplinaire de l'*Academic Word List* (AWL) de Coxhead [COX 00], tout en étant largement partagé par la plupart des disciplines, était assez inégalement réparti selon les familles disciplinaires. Cela ne nous parait toutefois pas être un argument qui remettrait en question l'intérêt du lexique transdisciplinaire. Comme nous l'avons illustré plus haut, le repérage des expressions inégalement partagées pourra précisément permettre de mieux comparer le fonctionnement linguistique et épistémologique des disciplines. Nous pensons par ailleurs que la mise en évidence de ce lexique constitue une entrée particulièrement pertinente pour réfléchir à la conceptualisation dans les disciplines. Le LST ne doit donc pas être conçu comme une liste « gravée dans le marbre », mais plutôt comme un matériau dont on observera la déclinaison dans les différentes disciplines et sous-genres,

en repérant les usages contextuels. Pour les applications didactiques d'aide à la rédaction scientifique (voir en particulier les chapitres 5, 6 et 8), il sera plus pertinent de travailler avec les apprenants sur des écrits en lien avec leur discipline de rattachement, qui leur seront plus facilement compréhensibles. Enfin, nous pensons qu'en nous limitant aux SHS, contrairement à l'AWL de Coxhead exploitée par Hyland et Tse [HYL 07], nous traitons une famille de champs disciplinaires plus homogène, où les pratiques scientifiques et méthodologiques présentent davantage de points communs[6].

Dans cet ouvrage, c'est ce lexique traversant des sciences humaines que nous traiterons et dont nous essaierons de cerner les contours.

I.2. Le lexique scientifique transdisciplinaire et les lexiques limitrophes

L'idée d'un lexique non spécifique aux disciplines, ou transversal, n'est pas nouvelle [HYL 07]. Des travaux précurseurs ont été proposés il y a plus d'une quarantaine d'années en français par Phal [PHA 72] avec le *Vocabulaire général d'orientation scientifique* (VGOS) dans la lignée du *Français fondamental* [GOU 56], ou en anglais dans le domaine du *subtechnical vocabulary* [AND 80]. Ce n'est évidemment que plus récemment que des entreprises lexicographiques de plus grande envergure, basées sur de grands corpus, ont pu être réalisées, à l'aide de méthodes de traitement automatique du langage et de linguistique de corpus.

Le terme de « lexique scientifique transdisciplinaire » a été forgé, à notre connaissance, par Patrick Drouin [DRO 07a]. Nous lui préférions auparavant l'appellation de « lexique scientifique des écrits scientifiques » [TUT 07a], qui mettait l'accent non seulement sur la dimension scientifique de ce lexique transversal, mais également sur les aspects discursifs, qui nous paraissent centraux. Le lexique partagé par les productions scientifiques est avant tout un lexique de genre, qui renvoie en effet non seulement aux procédures, démarches, objets scientifiques, mais aussi aux éléments d'argumentation, d'évaluation et de structuration du discours. Le métadiscours, brièvement défini comme le discours qui n'a pas une fonction référentielle, mais une fonction d'organisation du discours, d'évaluation et d'expression du point de vue [VAN 97], y occupe une place prépondérante. Ces éléments du métadiscours renvoient ainsi aux connecteurs, aux éléments de reformulation, marqueurs illocutoires, marqueurs épistémiques, modaux et évidentiels, marqueurs de points de vue, et marqueurs métatextuels [VAN 97]. Dans le discours scientifique, le guidage du lecteur, par

6. On notera toutefois que l'intersection entre le lexique transdisciplinaire propre aux sciences humaines et un lexique transdisciplinaire plus général, intégrant les « sciences dures » comme celui de Drouin [DRO 07a], semble assez importante, comme l'a montré l'étude de Hatier [HAT 16a].

exemple, s'exprime à travers des marqueurs de reformulation (1) ou à travers des marqueurs de renvoi intratextuel (2) :

> (1) Peut-on faire l'économie de la notion de sujet ? Pour le dire encore autrement, l'individu, porteur d'un sens subjectivement visé, est-il le niveau de réflexion adéquat de l'analyse sociologique ? [article, sociologie]

> (2) Qu'en est-il des autres formes de pc ? Comme on l'a dit plus haut, la définition en intension, tout d'abord, est incapable de remplir un rôle de modèle puisqu'elle ne représente rien. [article, linguistique]

Le lexique scientifique transdisciplinaire n'est donc pas le lexique de la science, mais bien celui du discours scientifique. En outre, il nous parait essentiel d'intégrer dans ce lexique les expressions polylexicales et les formules, comme celles qui sont soulignées en (1) et (2).

Si la notion de lexique transdisciplinaire est facile à appréhender, en définir les contours exacts reste malaisé. Il s'agit d'un objet souffrant d'un certain flou, car, contrairement à la terminologie, il n'est pas délimité par les disciplines. Par ailleurs, s'il est transdisciplinaire, les éléments qui le composent ne se distribuent pas également dans toutes les disciplines, comme on l'a vu plus haut. En outre, comme la plupart des unités lexicales fréquentes, et en particulier les mots simples, les mots du LST sont ambigus. Comme le soulignent Tremblay *et al.* (voir chapitre 7), seules certaines acceptions (ou « lexies » dans la terminologie des auteurs) du mot polysémique appartiendront au LST. Le mot *objet,* qui peut apparaitre comme un bon candidat au statut de LST, prend ainsi dans les articles de sciences humaines, selon le contexte, au moins quatre acceptions comme :

– élément terminologique de la linguistique (dans le sens de 'fonction grammaticale') :

> (3) Ce meilleur résultat des enfants sourds sur les entendants de 2P concernant le pronom sujet en modalité écrite, ainsi que leur faible taux de réussite concernant le pronom objet dans les deux modalités, seront repris dans la discussion. [article, linguistique]

– mot concret de la « langue générale » :

> (4) Plus loin se trouve un autre point de contrôle : deux militaires sont postés de chaque côté de la rue, ils inspectent à nouveau les objets transportés (photo 2). [article, géographie]

– nom abstrait du lexique scientifique transdisciplinaire dans le sens d'objectif' :

(5) Cette modification constitue l'objet de cet article qui s'attachera à décrypter la recomposition des équilibres fonctionnels portuaires induite par cette évolution institutionnelle. [article, géographie]

– nom abstrait du lexique scientifique transdisciplinaire avec le sens de 'élément, thème' :

(6) La dimension énonciative des productions textuelles, quant à elle, n'est pratiquement jamais considérée en tant que telle comme objet de traitement. [article, linguistique]

Si l'on sélectionne ainsi le mot *objet* comme appartenant au LST, il faudra en préciser les acceptions. C'est donc seulement en contexte que l'on peut déterminer le fonctionnement transdisciplinaire d'un mot.

En dehors de l'épineuse question de la polysémie, qui a été au cœur du projet TermITH (voir section I.5), distinguer les éléments de ce lexique des autres classes de lexèmes est souvent délicat [HAT 16a]. Les mots très fréquents de la « langue générale », tels que les mots grammaticaux (*de*, *la*, *sur*, *est*), sont évidemment faciles à écarter, ainsi que les mots du « français fondamental » [GOU 56], bien que certains d'entre eux puissent avoir des fonctions bien spécifiques dans le genre qui nous intéresse[7].

Plus complexe est le cas de ce qu'on a appelé le « lexique abstrait non spécialisé » [TUT 07a] ou « lexique abstrait général » [HAT 16a], qui intègre des unités lexicales abstraites (par exemple, *notion*, *argument*, *engendrer*, *pertinent*, *par conséquent*) surreprésentées dans les écrits scientifiques, mais qui sont aussi fréquentes dans les discours argumentatifs ou informatifs (la presse, par exemple). Comme on le verra dans les chapitres de cet ouvrage, le lexique transversal spécifique des écrits scientifiques comporte tout compte fait peu d'unités très spécifiques, comme *théorie*, *modèle*, *analyse statistique*, *méthode d'enquête*, mais de nombreux mots de l'abstraction, liés au raisonnement, au discours, à l'évaluation. Il apparait toutefois utile d'intégrer ce lexique abstrait général dans le LST du fait de son importance et de sa fréquence.

Une autre difficulté de délimitation apparait avec le « lexique des objets des sciences humaines et sociales » [HAT 16a, p. 77]. Ce champ lexical renvoie aux objets qui sont examinés de façon privilégiée par cette famille de discipline : l'homme, la ville, le pays,

7. C'est par exemple le cas du verbe *voir*, qui a une fonction évidentielle très marquée dans l'écrit scientifique et apparait dans des routines spécifiques (*comme on l'a vu*, *comme on peut le voir*…) [GRO 10b].

la région, les classes sociales… Comme nous le verrons plus loin, les techniques lexicométriques employées pour mettre en évidence le LST extraient nécessairement ce champ lexical fréquent et commun à de nombreuses disciplines des SHS. Dans l'approche du LST développée dans le cadre du projet TermITH, ce lexique a toutefois été écarté, car il n'est pas directement lié au discours et à la démarche scientifique. Il serait néanmoins tout à fait pertinent d'envisager son étude de façon plus systématique, en particulier dans une perspective comparative des écrits de SHS.

Enfin, certains éléments du LST peuvent être très proches de la terminologie des disciplines, du fait de leur polysémie ou parce qu'ils sont fréquemment intégrés dans les termes complexes (voir section I.5 et chapitre 4).

À titre d'exemple, nous présentons dans l'encadré I.1 un extrait d'article annoté comportant les mots du LST (en italique, à partir de la liste proposée en annexe) et les termes disciplinaires (en bordeaux). Les autres types de mots ne sont pas traités.

La *fonction* de notre analyseur est d'*identifier* des relations de dépendances entre mots et d'extraire d'un corpus des syntagmes (verbaux, nominaux, adjectivaux). Le *résultat* de l'*analyse* se *présente* sous la *forme* d'un réseau de dépendance, dans lequel chaque syntagme extrait est *relié* à sa tête et à son expansion syntaxiques (*figure 1*). Ces relations de dépendance *permettent* d'*effectuer* automatiquement des regroupements distributionnels : *par exemple* la liste de tous les compléments de tel verbe ou la liste des adjectifs modifieurs de tel nom, qui *constituent* des amorces de classes sémantiques. À titre d'*illustration*, l'*analyse* du réseau *présenté* sur la *figure* 1 *suggère* un regroupement des noms « alluvion », « sable » et « lave » qui sont tous les trois arguments des verbes « disparaitre sous » et « creuser dans ».

Encadré I.1. *Les différents types de lexiques*[8]

Cet exemple appelle plusieurs commentaires. Tout d'abord, on observe une forte proportion du LST, tel qu'on l'a défini, parmi les mots du texte (18 mots sur 124, soit 14,5 % des mots) alors que les termes représentent eux 25 % du total. Par ailleurs, comme mentionné plus haut, peu d'unités lexicales apparaissent véritablement spécifiques des écrits scientifiques (hormis *figure*, *analyser*). C'est la récurrence des lexèmes du LST et les cooccurrences les associant (*la fonction est d'identifier*, *résultat de l'analyse*…) qui donnent à l'extrait une tonalité scientifique. Enfin, il peut se révéler délicat de départager terminologie et LST. Deux mots se montrent ici particulièrement complexes à cet égard : *mots* et *corpus*. Bien que ces deux mots soient intégrés dans la liste du LST, ils ont été considérés dans l'extrait comme des termes spécialisés du domaine de la linguistique, car ils prennent dans ce contexte une acception spécialisée.

8. Bourigault D., Fabre C., « Approche linguistique pour l'analyse syntaxique de corpus », *Cahiers de grammaire*, n° 25, p. 131-151, 2000.

La délimitation du LST apparait donc parfois délicate. Dans le cadre du projet TermITH, le recours à une classification sémantique, à l'instar de Pecman [PEC 04], nous a toutefois aidés à organiser ce lexique et à en circonscrire les limites (voir chapitres 1, 5 et 6).

I.3. Les ressources du LST

Le lexique transversal des écrits scientifiques fait depuis une quarantaine d'années l'objet de traitements lexicographiques. Nous indiquerons dans cette section les principales ressources du français – et plus ponctuellement, de l'anglais – sans prétendre à l'exhaustivité. Nous privilégions ici les ressources qui ont eu pour objectif d'élaborer des inventaires lexicaux, parfois accompagnés de traitements sémantiques et syntaxiques, plutôt que les travaux de nature purement théorique (qui seront davantage abordés dans les chapitres de l'ouvrage).

Les ressources du LST présentent des points communs. D'une part, les méthodes d'extraction et de sélection du LST exploitent, même pour les plus anciennes d'entre elles comme le *Vocabulaire général d'orientation scientifique* (VGOS) [PHA 72], des corpus et des mesures lexicométriques communes.

– La mesure de fréquence, calculée sur des formes ou des lemmes, vérifie que les mots extraits sont productifs.

– La mesure de dispersion entre les sous-corpus des disciplines permet d'assurer le caractère transversal du lexique et d'écarter la terminologie propre à une discipline.

– La mesure de spécificité, souvent calculée par opposition à un corpus de contraste, sélectionne les mots qui sont particuliers à ce corpus et écarte les mots banals, relevant de la langue générale.

D'autre part, en ce qui concerne les mots simples, peu de ressources présentent, comme le signalent Tremblay et ses collègues dans ce volume (chapitre 7), des traitements sémantiques.

I.3.1. *Les ressources lexicales de mots simples*

Si l'extraction des mots simples du LST à l'aide de techniques lexicométriques est assez simple, la sélection et le traitement de ce lexique posent paradoxalement des problèmes plus complexes que pour les expressions polylexicales. La principale

difficulté, déjà mentionnée à la section précédente, est liée à la polysémie de ces mots. D'une part, si un mot est fréquent dans un ensemble de disciplines, il faut vérifier que les acceptions transdisciplinaires de ce mot sont aussi partagées, ce qui suppose donc d'opérer un tri manuel. D'autre part, pour les diverses applications du LST, une mise en contexte du lexique dans des phrases d'exemples, ou un traitement sémantique, s'avèrera indispensable.

Pour le français, la ressource la plus ancienne et la plus connue est sans conteste le VGOS (*Vocabulaire général d'orientation scientifique*) de Phal [PHA 72] qui, dans la lignée du *Français fondamental* de Gougenheim [GOU 56], a proposé de constituer un lexique fondamental de la science, principalement pour l'enseignement du français langue étrangère. L'objectif de Phal et de ses collègues était de mettre en évidence le lexique scientifique récurrent et spécifique, essentiellement dans des manuels de sciences (physique, mathématiques, chimie, sciences naturelles) du collège au lycée, et de l'extraire à l'aide de techniques lexicométriques. Dans la continuité de cette étude, Lamria Chetouani [CHE 97] a proposé une version légèrement différente de ce lexique, qu'elle a appelé le « Vocabulaire général d'enseignement scientifique » (VGES), basé sur des discours scientifiques oraux dans des classes de lycées incluant essentiellement des sciences expérimentales (physique) et quelques cours de droit en sciences sociales. Le VGOS et le VGES sont avant tout destinés au contexte scolaire, et sont construits à partir de discours scientifiques à fonction pédagogique (manuels et cours) et non à partir d'articles de recherche. On y retrouve donc de nombreux lexèmes propres aux sciences dures (par exemple : *bissectrice, calcaire, chlore*), mais moins de lexèmes renvoyant à la rhétorique argumentative ou à des opérations d'analyse (par exemple, les mots *analyser, rejeter, point de vue* présents dans notre LST y sont absents)[9]. En outre, si certaines acceptions sont distinguées, la liste reste assez brute et ne propose pas de traitement sémantique pour les mots extraits.

Des études plus récentes, exploitant le traitement automatique et la linguistique de corpus, ont prolongé le travail pionnier de Phal dans les années 2000. L'intérêt des chercheurs s'est alors davantage tourné vers les écrits scientifiques (plutôt que vers les manuels à fonction didactique). Les objectifs étaient de constituer des ressources lexicales visant des applications pédagogiques, mais aussi linguistiques et terminologiques. Drouin [DRO 07a] a proposé une extraction en français (et en anglais, par alignement) du LST à partir d'un corpus de thèses de 2,3 millions de mots dans 9 disciplines des sciences humaines et sciences « dures ». Utilisant des mesures de spécificité (avec *Le Monde* comme corpus de contraste) et de répartition (50 % des tranches), il a extrait

9. On retrouve en revanche plusieurs champs lexicaux communs au VGOS et au LST TermITH, comme les mesures, le temps, les qualités, certaines procédures.

1 113 mots du LST sous forme lemmatisée des quatre grandes catégories ouvertes (nom, adjectif, verbe, adverbe). L'intersection de ce lexique avec le VGOS s'est révélée assez limitée, en particulier pour les verbes, probablement du fait de la différence de corpus (par exemple, les verbes *découler, inspirer* ou *éclairer* présents chez Drouin sont absents du VGOS). Drouin a également proposé un repérage des collocations transdisciplinaires de type Verbe-Nom (voir section I.3.2). Enfin, dans une version plus récente du LST, accessible en ligne[10], Drouin et son équipe ont proposé un traitement des acceptions, associées aux correspondances en anglais, à la granularité assez fine. La figure I.2 illustre le traitement proposé pour le mot du LST *objet*, qui présente 4 acceptions, et les liens vers les termes anglais correspondants.

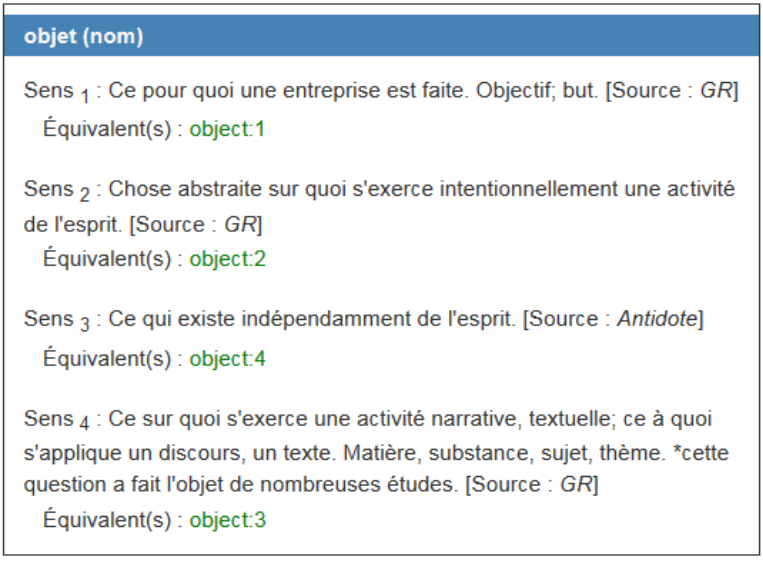

Figure I.2. *Le traitement du mot objet dans le LST de l'OLST*

Dans notre équipe au LIDILEM, nous avons parallèlement proposé d'élaborer une première liste du LST [TUT 10a][11], à partir d'un corpus diversifié d'écrits scientifiques de 2 millions de mots (rapports, thèses et articles en linguistique, économie et médecine), à l'aide d'une méthode lexicométrique simple (lexèmes apparaissant plus de 15 fois dans les trois disciplines). Un sous-ensemble de cette liste a reçu un traitement sémantique,

10. Le LST élaboré par Drouin est interrogeable à l'adresse : http://olst.ling.umontreal.ca/?page_id=1511.

11. La liste est disponible à l'adresse : https://scientext.hypotheses.org/lexique-transdisciplinaire-v1.

mais c'est essentiellement le travail mené dans le cadre du projet TerMITH [HAT 16a], présenté dans cet ouvrage (voir section I.5), qui a permis de proposer une méthodologie plus rigoureuse et un traitement plus adéquat.

Enfin, parmi les travaux récents, signalons également la liste du « Vocabulaire savant de base » [DAS 10] dans la perspective de l'indexation automatique. Ce vocabulaire nominal de 550 éléments, extraits par des méthodes de traitement automatique des langues (TAL) d'un corpus de textes scientifiques, s'articule avec le vocabulaire technique spécialisé pour constituer des entrées complexes pour l'indexation.

Les travaux sur le LST en anglais sont très nombreux et nous mentionnerons principalement l'*Academic Word List* (AWL)[12] de Coxhead [COX 00], très largement exploitée dans les travaux autour de l'*English for Academic Purposes*. L'AWL a été constituée à l'aide d'un corpus diversifié de 2,8 millions de mots dans 28 domaines des sciences humaines, sciences dures et droit. Ont été sélectionnés les mots fréquents (plus de 100 occurrences apparaissant dans 4 des 8 grands domaines), n'appartenant pas aux 2 000 mots les plus fréquents de l'anglais. La liste est organisée par tranches de fréquence et contient des familles morphologiques comme *available/availability/ unavailable*. Dans la plus récente *Academic Keyword List* (AKL), Paquot plaide quant à elle pour l'intégration de certains mots fréquents, écartés de l'AWL [PAQ 10].

I.3.2. *Les ressources phraséologiques du LST*

La phraséologie du lexique scientifique transdisciplinaire a suscité beaucoup de travaux, particulièrement dans le champ très dynamique de l'*English for Academic Purposes*. La dimension phraséologique du lexique du genre scientifique apparait essentielle, à plusieurs titres. Tout d'abord, c'est avant tout à travers la combinatoire lexicale que l'on peut observer le fonctionnement sémantique du LST. Le mot *terme* dévoile ainsi son ambigüité dans ses cooccurrences : (1) *au terme de*, *en arriver au terme* (2) *en d'autres termes*, *définir un terme*… Par ailleurs, les expressions polylexicales, moins ambigües, sont plus faciles à traiter sur les plans sémantique et pragmatique. Plusieurs ressources phraséologiques comportent d'ailleurs des informations sémantiques et pragmatiques qui faisaient défaut aux ressources de mots simples. Enfin, les applications didactiques (entre autres) sont friandes de ces expressions préfabriquées qui permettent de comprendre le fonctionnement rhétorique du genre. Sans toutefois proposer ici un panorama complet des travaux sur la phraséologie scientifique transdisciplinaire, nous rappellerons quelques jalons importants du domaine.

12. Disponible à l'adresse : https://www.victoria.ac.nz/lals/resources/academicwordlist.

Si Phal avait souligné l'importance de la dimension syntagmatique dans le VGOS, qui intègre d'ailleurs quelques locutions, le premier inventaire phraséologique d'envergure pour le français est à notre connaissance celui de Pecman [PEC 04]. Dans son travail de thèse, elle a exploité un corpus bilingue de sciences dures (chimie, physique et biologie) permettant d'extraire et de modéliser 2 000 unités phraséologiques, en utilisant un étiquetage notionnel autour de quatre grandes sphères abstraites : la scientificité, l'universalité, mais aussi la modalité et la discursivité, qui intègrent des aspects discursifs et rhétoriques. Les notions, par exemple |EXPÉRIMENTATION| ou |RÉSULTAT|, servent à organiser le classement phraséologique, sous forme de schémas collocationnels. Cette démarche originale n'a toutefois donné lieu qu'à des inventaires partiels et la liste complète des unités phraséologiques n'a malheureusement pas été publiée.

À la suite de ses travaux sur les mots simples (voir section I.3.1), à partir du même corpus, Drouin [DRO 07a] a également proposé un traitement des collocations de type Verbe-Nom, en exploitant des mesures statistiques, mais ce premier inventaire n'a pas encore fait l'objet d'un traitement sémantique. Du côté de la modélisation des expressions, le mémoire de master de Pouliot [POU 12] est une réflexion linguistique et didactique intéressante sur la présentation, dans un dictionnaire destiné aux apprenants, des collocations organisées sous la forme de regroupements sémantique et syntaxique, mais la description se limite à quelques entrées nominales.

Les travaux sur la phraséologie scientifique sont très nombreux pour l'anglais et plusieurs ressources phraséologiques importantes ont récemment vu le jour. Parmi les études théoriques majeures sur la question, on mentionnera l'étude de Gledhill [GLE 00] sur les collocations scientifiques, au sens de Sinclair [SIN 91], en particulier pour la notion de *generic collocation*, qui renvoie aux collocations scientifiques traversantes. Biber et ses collègues [BIB 04] ont proposé quant à eux des classifications fonctionnelles des *lexical bundles*, suites de mots récurrentes qui n'ont pas toujours un statut phraséologique clair, propres au discours académique oral et écrit.

Se basant sur la classification de Biber, Simpson-Vlach et Ellis [SIM 10] ont développé une ressource ambitieuse pour l'anglais, qui se présente comme l'équivalent phraséologique de l'AWL de Coxhead, l'*Academic Fomulas List* (AFL)[13], et qui recense les séquences formulaires propres aux discours oraux et écrits des discours académiques. Ces expressions sont extraites à partir d'un corpus oral et écrit de 4 millions de mots (*Michigan corpus*, extraits du *British National Corpus* (BNC), en repérant les expressions fréquentes, réparties dans plusieurs tranches de textes et spécifiques à ce

13. La liste est consultable à l'adresse : http://www.eapfoundation.com/vocab/academic/afl/.

genre. Les séquences (de 3 à 5 mots uniquement) sont sélectionnées à l'aide de mesures statistiques, qui ont été considérées comme pertinentes à partir d'un ensemble d'expressions examinées (*formula teaching worth*) par des enseignants. Six cent sept séquences sont ainsi extraites, dont une grande partie est caractérisée sur le plan pragmatique et sémantique à l'aide des trois grandes classes définies par Biber [BIB 04] : expressions à fonction référentielle, interpersonnelle (*stance*) ou à fonction discursive, qui sont ici affinées. L'idée d'exploiter les mesures associées à l'intérêt pédagogique et le classement sémantique proposé donnent à cette ressource un réel intérêt malgré certaines limites : l'exclusion des formules binaires, l'inclusion de segments peu pertinents extraits automatiquement (comme *in other words the*), et l'absence de typologie linguistique de ces expressions, qui mélangent collocations, expressions figées et motifs compositionnels.

Plus récemment, des lexicographes de l'éditeur Pearson, Kirsten Ackermann et Yu-Hua Chen [ACK 13] ont élaboré une ressource lexicographique très complémentaire de l'AFL, l'*Academic Collocation List* (ACL). Les expressions traitées sont ici exclusivement des collocations lexicales (par exemple, *abstract concept* ou *fully integrated*), entendues comme des associations binaires privilégiées. La ressource a été constituée à partir d'un corpus de 25,6 millions de mots, comprenant des articles de recherche et des ouvrages dans les principaux domaines de recherche. Les associations sont extraites à partir de techniques de TAL et de mesures statistiques d'association[14] (information mutuelle et t-score), en exploitant également la fréquence et la répartition. Mais ce qui caractérise surtout cette ressource – et en garantit la qualité – est sa validation par un ensemble de 6 experts qui en ont déterminé l'intérêt linguistique et pédagogique. Au terme du processus, l'ACL[15] contient 2 468 éléments relevant de structures syntaxiques variées, intégrant une information de fréquence et quelques éléments contextuels. La ressource ne comporte toutefois, à notre connaissance, aucun traitement sémantique ni exemple.

En bref, les dernières années ont connu un réel développement de ressources lexicales portant sur le lexique transversal des écrits scientifiques. Si les techniques de linguistique de corpus apparaissent centrales, un traitement manuel demeure indispensable, pour sélectionner les unités lexicales pertinentes, distinguer les acceptions et proposer des traitements sémantiques et pragmatiques.

14. Sans utiliser toutefois la syntaxe, ce qui conduit à des post-traitements un peu fastidieux, que nous évitons dans notre approche (voir chapitre 3).

15. Disponible à l'adresse : https://pearsonpte.com/organizations/researchers/academic-collocation-list/.

I.4. Le LST développé dans le cadre du projet TermITH

Dans le cadre du projet TermITH (voir section I.5 et chapitre 4), le LIDILEM a été chargé d'élaborer une liste du lexique scientifique transdisciplinaire pour faciliter l'indexation des textes de sciences humaines. Différentes facettes de ce travail sont présentées dans cet ouvrage : l'extraction et la modélisation des noms (chapitre 1), le traitement des verbes (chapitre 5) et des adverbes (chapitre 6), l'extraction et le traitement des expressions polylexicales (chapitre 3). Comme les traitements proposés sont résumés dans les travaux de Hatier [HAT 16a, HAT 16b] et dans le chapitre 1 de cet ouvrage, nous n'entrerons pas ici dans les détails, mais mentionnerons simplement les grandes étapes de l'extraction, de la sélection et de la modélisation du LST.

La méthode d'élaboration du LST – dans les étapes d'extraction, sélection et modélisation – a été largement fondée sur des ressources textuelles qui se devaient d'être caractéristiques du genre. À cette fin, un corpus représentatif des écrits de sciences humaines a été constitué, le corpus Transdisciplinaire TermITH, tout d'abord dans le cadre du travail de thèse de Tran [TRA 14], complété dans un deuxième temps par Hatier [HAT 16a]. Ce corpus a servi à la constitution des ressources lexicales, mais a aussi été exploité dans de nombreux travaux de notre équipe.

Ce corpus, décrit plus en détail dans [TRA 14] et [HAT 16a, p. 56-59], comporte 500 articles répartis dans 10 disciplines de sciences humaines, tirés de revues de référence (voir tableau I.1). Le corpus a été annoté sur le plan structurel et analysé syntaxiquement.

Discipline	Nombre d'articles	Nombre de mots
Anthropologie	50	493 988
Économie	50	417 944
Géographie	50	400 533
Histoire	50	773 170
Linguistique	50	425 952
Psychologie	50	417 846
Sciences de l'éducation	50	417 069
Sciences de l'information	50	399 007
Sciences politiques	50	548 222
Sociologie	50	540 630
Total	**500**	**4 834 361**

Tableau I.1. *Composition du corpus Transdisciplinaire TermITH ([HAT 16a, p. 58])*

Pour exploiter des mesures de spécificité (utilisées dans notre lexique, à l'instar de Drouin ou de Paquot, voir section I.3.1), un corpus de contraste diversifié de 119 millions de mots a été élaboré par Hatier [HAT 16a, p. 60-63]. Ce corpus, intégrant des textes écrits de fiction, de presse nationale et de presse régionale, ainsi que des textes oraux transcrits (sous-titres de films, émissions de radio), a également été annoté syntaxiquement.

La méthode d'extraction et de sélection des mots simples est décrite pour les noms dans le chapitre 1 et dans [HAT 16a, 16b]. Elle exploite des traitements lexicométriques automatiques et un processus de sélection manuel. Les mesures employées ont permis d'extraire : (a) les mots spécifiques (ayant un ratio supérieur au corpus de contraste), (b) répartis dans au moins 4 des 10 disciplines et dans 40 des 100 sections du texte, (c) n'appartenant pas à des expressions figées du type *mise en œuvre*. Comme cette première liste comportait un peu de bruit et quelques erreurs d'étiquetage automatique, une méthode de filtrage manuel à l'aide de juges a été mise en place pour sélectionner, à l'aide d'exemples, les éléments pertinents (voir chapitre 1).

Comme il a été rappelé plus haut, une simple liste des formes était de peu d'intérêt pour les applications envisagées. Un ensemble de traitements linguistiques (voir tableau I.2 pour quelques exemples), s'appuyant sur le corpus de référence, a alors été effectué (voir chapitres 1, 5 et 6) :

– repérage des acceptions du LST. Par exemple, dans le cadre du LST, le nom *conclusion* recevra deux acceptions, selon qu'il prend le sens de section d'un article ou de relation logique ;

– classement sémantique des acceptions du LST dans de grandes classes sémantiques et sous-classes sémantiques, dont les définitions sont présentées en annexe 2. Par exemple, l'adjectif *strict* et l'adverbe *strictement* recevront tous deux les mêmes étiquettes sémantiques ('modalité') et la même sous-classe ('restriction'). Ces classes sémantiques ont été élaborées à partir de critères rigoureux, en partie à l'aide de techniques de TAL ;

– les acceptions sont enfin accompagnées de gloses, souvent tirées du *Dictionnaire électronique des mots* [DUB 10] ou des *Verbes français* [DUB 97], parfois complétées par d'autres ressources, et illustrées par des exemples.

La liste des mots du LST, accompagnés des classes sémantiques et des gloses, apparait en annexe 1 de cet ouvrage. Elle comprend également des phrasèmes complets ou locutions (voir chapitre 3), par exemple *point de vue* ou *faire face*. La liste des

exemples de mots simples et locutions, ainsi qu'une liste de collocations (voir traitement au chapitre 3) accompagnées d'exemples, est accessible en ligne[16].

Acception	Lemme	Partie du discours	Classe sémantique	Sous-classe sémantique	Glose
conclusion-1	conclusion	nom	communication_ support	section	section qui termine un article, un ouvrage
conclusion-2	conclusion	nom	relation	implication	déduction, conséquence
strict	strict	adjectif	modalité	restriction	limité
strictement	strictement	adverbe	modalité	restriction	rigou- reusement

Tableau I.2. *Exemples de traitement du LST ([HAT 16b])*

I.5. Le projet TermITH

Une bonne partie des développements récents dans l'équipe LIDILEM autour du LST a eu pour cadre le projet de recherche TermITH (terminologie et indexation de textes en sciences humaines)[17]. Ce projet a poursuivi un objectif principal d'indexation de textes scientifiques produits dans les disciplines des sciences humaines, par exploitation de la terminologie présente dans ces textes. Cet appui sur la terminologie mobilisée dans les textes scientifiques a impliqué de résoudre plusieurs questions, qui ont alors constitué autant de sous-objectifs du projet :

– extraire automatiquement les termes, donc identifier des termes potentiels (que l'on nommera alors *candidats-termes*) et les lister ;

16. Disponible à l'adresse : http://lidilem.u-grenoble3.fr/ressources/corpus-du-labo/article/lexique-scientifique.

17. Projet ayant reçu un financement ANR dans le programme CONTINT (référence ANR-12-CORD-0029) et réunissant six laboratoires de recherche français : l'ATILF (Analyse et traitement informatisé de la langue française), l'InIST (Institut de l'information scientifique et technique), le LORIA (Laboratoire lorrain de recherche en informatique et ses applications), le LINA (Laboratoire d'informatique de Nantes Atlantique), le LIDILEM (Laboratoire de linguistique et didactique des langues étrangères et maternelles) et l'INRIA (Institut national de recherche en informatique et en automatique).

– désambigüiser et valider les termes, ce qui signifie d'écarter certains des candidats-termes qui sont présentés par le système d'extraction. Cela revient à déterminer des critères et des procédures permettant de faire, de façon aussi automatique que possible, la distinction entre des unités qui, au final, manifestent des caractéristiques terminologiques et celles qui ne relèvent pas de la terminologie du domaine.

L'une des difficultés à surmonter réside dans l'analogie formelle entre termes et non-termes. Des extraits d'articles de recherche en linguistique mettront en évidence ces ressemblances formelles :

> (7) En 1865, Baillarger met en évidence une activité langagière qui, imputable à un locuteur, lui échappe néanmoins. [article, linguistique]
>
> (8) *Les enfants* jouent *sur la terrasse* est l'exemple qui fait apparaitre l'un des sens réputés de base du verbe *jouer*, à savoir 'se livrer à une activité ludique'. [article, linguistique]

On voit dans les deux cas le nom *activité* suivi d'un adjectif, mais le statut de ces groupes est différent : *activité langagière* est un terme complexe de linguistique, qui renvoie à un concept usuel du champ, alors que *activité ludique* est une association libre d'un nom et d'un adjectif, ne renvoyant à aucun concept particulier de la discipline. Une extraction automatique, en particulier en vue de déterminer les descripteurs appropriés d'un texte scientifique pour son indexation, doit alors mettre en œuvre une analyse des candidats-termes et de leur environnement afin de désambigüiser les diverses occurrences.

Le projet TermITH a travaillé sur une approche novatrice de cette analyse, en croisant les candidats-termes extraits avec, d'une part, une liste d'unités du lexique scientifique transdisciplinaire et, d'autre part, des ressources terminologiques déjà disponibles, et ce afin d'isoler les termes propres au domaine étudié [BIL 14, JACQ 13]. Cette approche présente deux intérêts principaux :

– la désambigüisation et le croisement avec les lexiques transdisciplinaires et les ressources terminologiques disponibles réduisent l'effort humain de vérification des termes (ou descripteurs) proposés et l'analyse manuelle du contenu des documents ;

– la mise à jour des ressources terminologiques peut être automatisée.

La désambigüisation est nécessitée par une particularité des textes scientifiques des disciplines des SHS par rapport à d'autres champs scientifiques, illustrée par les extraits (7) et (8) ci-dessus et bien explorée dans le chapitre 4 : les termes sont fréquemment des

unités lexicales qui existent dans la langue avant même d'être des termes que les disciplines ont spécialisées, soit en les associant à des adjectifs ou à des constructions particulières, soit en spécifiant leur sens. Par exemple, *production verbale* est un terme de linguistique, *production agricole* ou *production industrielle* sont des termes d'économie, et dans le même temps, *production* est défini dans un dictionnaire tel que le TLFi[18] comme « action d'engendrer, de faire exister ; le fait ou la manière de se produire, de prendre naissance ». Cette dernière signification est présente dans l'extrait (9) et c'est précisément celle qui est associée au nom *production* en tant qu'unité du LST :

> (9) Une partie considérable de l'activité de production de titres, slogans ou formes à succès repose sur cette sorte de polyphonie par laquelle les énoncés s'interpellent. [article, linguistique]

Désambigüiser signifie alors, pour les candidats-termes extraits, décider s'ils relèvent bien de la terminologie du domaine, ou plutôt du LST, ou encore d'aucune de ces catégories.

La réalisation du projet a ainsi conduit à entreprendre, approfondir, compléter des recherches linguistiques dans deux directions principales, dont on trouvera de nombreuses traces dans les chapitres qui composent l'ouvrage. Comme on l'a explicité dans la section I.4, l'inventaire du lexique scientifique transdisciplinaire, dont les contours avaient déjà été établis depuis plusieurs années, devait être complété et affiné. Puisque ce lexique était destiné à entrer dans les procédures de désambigüisation, il était absolument nécessaire d'en fournir une liste dans toutes les catégories lexicales et grammaticales dans lesquelles ses unités se manifestent, et pour les unités aussi bien simples que complexes. Divers travaux ont donc consisté à extraire ces unités des textes scientifiques eux-mêmes, en mêlant phases automatisées et validation manuelle, selon la démarche qui a été exposée section I.4 et qui est explicitée dans le chapitre 1 de l'ouvrage. En parallèle, puisque le projet visait une indexation des textes par leur contenu [BOUG 15], il ne s'agissait pas seulement d'analyser des unités, candidats-termes extraits de leur environnement, mais de comprendre comment le contexte de ces unités pouvait être mis à contribution pour la désambigüisation de ces unités. Pour mieux comprendre cette dernière problématique, là encore, des extraits d'articles scientifiques illustreront les questions à résoudre. Y sont mises en valeur les unités du LST et coloriées en bleu les unités terminologiques :

18. *Trésor de la langue française informatisé*, disponible à l'adresse : http://atilf.atilf.fr/tlfi.htm.

(10) Je voudrais dans ce qui suit examiner brièvement le *traitement* des constructions de type *En tout(e) N.* [article, linguistique]

(11) La manière dont Ducrot s'efforce de *définir* la signification linguistique est encore autre chose. [article, linguistique]

Il apparait clairement dans ces deux extraits une relation de rection syntaxique entre les unités du LST, *traitement* et *définir*, et les termes constructions et signification linguistique. Une seconde direction de recherches a ainsi exploré le champ des relations qui s'établissent, dans les textes, entre les termes et les unités du lexique scientifique transdisciplinaire. Un échantillon des travaux menés dans ce cadre est donné par le chapitre 4.

Quoique, pour l'enrichissement des terminologies existantes, le projet TermITH ait porté au premier chef sur les termes, leur extraction et donc leur analyse, à la fois en contexte et hors contexte, les unités du LST y ont joué un rôle non négligeable, pour deux raisons :

i) les caractéristiques formelles des termes des SHS qui, on l'a vu, peuvent être bâtis sur les mêmes formes que celles qui appartiennent au LST ;

ii) la participation de ces unités du LST à l'organisation, à la fois sémantique et rhétorique, du discours scientifique.

Les chapitres de cet ouvrage, tour à tour, éclairent ces rôles discursifs des unités, explorent les relations entre LST et terminologie et proposent des pistes pour l'enseignement de ces unités aux étudiants, allophones et/ou futurs chercheurs.

I.6. Présentation de l'ouvrage

Outre cette introduction, l'ouvrage s'organise en deux parties, chacune comportant quatre chapitres qui couvrent différents aspects d'une même thématique. La première partie rassemble des études descriptives du LST, la seconde se centre sur les aspects didactiques.

Le chapitre 1, rédigé par Sylvain Hatier, explicite la construction d'une ressource lexicale en français, exploitable aussi bien pour des traitements automatiques de la langue – tels que ceux qui ont été menés dans le projet TermITH (voir section I.5) –, que pour des applications didactiques, dont les chapitres 5 et 6 de la seconde partie donneront des exemples qui s'appuient sur cette ressource. La ressource non seulement inventorie les unités du LST, acquises par une procédure mixte, automatisée et manuelle, à partir d'un corpus d'articles de recherche en sciences humaines et sociales, mais propose en

outre un classement sémantique de ces unités. Avec ce classement, Sylvain Hatier et les collègues qui participent aux travaux (voir aussi le chapitre 3) vont bien au-delà des réalisations de LST précédentes, évoquées section I.3. C'est l'organisation sémantico-pragmatique du travail de recherche qui est dévoilée à travers des classes telles que « Observation/#collecte » (verbes tels que *collecter, regrouper, rassembler*), « Processus_humain/#usage » (verbes tels que *employer, mettre en œuvre*). Une force de cette ressource est d'être ancrée sur les écrits de recherche effectivement produits dans le champ. Il ne s'agit donc pas d'une description intuitive, mais bien d'une forme de modélisation de ce que les textes eux-mêmes donnent à voir, dans une tradition maintenant bien établie en TAL et en linguistique, de construction de ressources à partir de corpus (par exemple [MES 10]). Une autre force tient au classement sémantique des unités et au niveau d'abstraction supplémentaire qu'il rend possible. Ce classement permet des observations non sur les unités, mais sur des classes plus abstraites, ce qui fait émerger de nouvelles régularités quant à la façon dont ces classes d'unités participent à la construction du discours scientifique, ce qui est développé notamment dans le chapitre 3 par Agnès Tutin.

Le chapitre 2, dont l'auteur est Francis Grossmann, aborde précisément la façon dont le discours scientifique intègre des formes adverbiales du champ lexical de la généralisation pour marquer certaines opérations logico-cognitives. Il analyse très précisément le fonctionnement discursif d'un type d'adverbes qu'il nomme *adverbes d'habitude*, en raison du fait qu'ils permettent de signifier que le procès décrit se déroule « habituellement », « en général », par exemple en (12) :

(12) Deux axes, corrélés entre eux, sont généralement cités. [thèse, TAL]

L'étude de Grossmann peut être mise en relation avec celle présentée par Tran au chapitre 6. Tran présente une mise en ordre typologique des adverbiaux relevant du LST et donc présents dans l'écrit scientifique ; Grossmann « zoome » sur une sous-catégorie précise et en analyse très finement le fonctionnement rhétorique et discursif. Après avoir situé son étude et ses objets dans le champ des analyses de sémantique lexicale, et avoir rappelé les diverses facettes de la notion de généralisation, Grossmann met en évidence trois grandes sphères sémantiques pour les adverbiaux étudiés : un sens d'habitude, un sens métalinguistique, un sens d'approximation. Les effets sémantiques et rhétoriques des adverbiaux dans chacune de ces sphères sont ensuite soigneusement distingués et décrits. L'étude de Grossmann est ainsi de celles qui peuvent nourrir les dispositifs d'enseignement de l'écriture scientifique, tels ceux qui sont évoqués dans les chapitres 6 et 8 (ou aussi dans [BOC 15]), en ce qu'elle permet une compréhension fine des ressorts linguistiques de la modalisation et de la généralisation dans la production des textes scientifiques, articles, mémoires ou autres écrits.

En effet, l'écriture scientifique, sans être aussi normée dans les disciplines de sciences humaines et sociales que dans les disciplines plus expérimentales, n'est pas totalement libre et laissée au gré de la créativité de l'auteur. Elle s'appuie sur une phraséologie et sur des « routines », au sens de « séquences lexico-syntaxiques "prêtes à écrire", étroitement associées à des pratiques discursives spécifiques, liées à des communautés bien délimitées » [SIT 16], qui participent à faire de l'article de recherche un genre à part entière. Le chapitre 3, sous la plume de Tutin, explore cette phraséologie et ces routines, à la fois sous l'angle de leur inventaire, basé sur des méthodes d'extraction similaires à celles qui sont exposées dans le chapitre 1, et sous celui des traitements linguistiques qui permettent de les constituer en ressource lexicographique. Tutin s'intéresse successivement à trois types d'expressions :

1) les locutions, telles *point de vue*, dont le sens n'est pas compositionnel et qui présentent un fort degré de figement syntaxique ;

2) les collocations, telles *jouer un rôle*, dont le sens est, totalement ou partiellement, compositionnel ;

3) les routines, telles *comme nous l'avons évoqué précédemment,* qui sont de l'ordre de la formule et dont le sens est pleinement compositionnel.

Chacun de ces types requiert une méthodologie adaptée, autant pour être mis au jour à partir de corpus que pour être décrit et caractérisé dans une ressource lexicographique. Tutin s'est appuyée pour l'extraction sur le corpus transdisciplinaire constitué dans le cadre du projet TermITH (voir section I.5), analysé en dépendances syntaxiques, ce qui a permis de faire intervenir des arbres syntaxiques dans l'analyse, et elle mobilise la catégorisation du LST présentée au chapitre 1. Son étude montre comment capter l'interaction lexique-discours dans les textes scientifiques.

C'est une autre interaction, celle des unités du LST avec la terminologie des disciplines, qui est analysée dans le chapitre 4, rédigé par Jacquey, Kister, Marcon et Barreaux. Dans le cadre du projet TermITH, l'inventaire du LST était en premier lieu destiné à fournir une forme de liste d'exclusion pour l'extraction automatique des termes. En effet, les techniques d'extraction proposaient régulièrement comme candidats-termes des groupes nominaux complexes tels que *étude des prises de risques* ou *traitement de critiques de danse,* dans lesquels le premier nom (ici *étude* ou *traitement*) est une unité du LST. L'identifier comme telle permettrait ainsi de réduire le nombre de candidats-termes et d'alléger la tâche. Dans le même temps, la récurrence de telles associations ainsi que d'autres observations ont conduit les auteurs à s'interroger plus particulièrement sur les relations entre LST et terminologie. Les auteurs abordent ces relations sur deux plans : 1) la présence des unités du LST dans les syntagmes terminologiques de type Nom-Adj*,* par exemple *analyse syntaxique* ; 2) la dépendance syntaxique, en contexte, d'unités du LST et de candidats-termes, par exemple *analyser le*

langage intérieur. Leur étude, essentiellement quantitative, livre les diverses interprétations possibles pour les candidats-termes de type Nom-Adj, c'est-à-dire le choix, pour chacun des deux composants, entre trois types d'acceptions : un sens terminologique, lié à la discipline ; un sens transdisciplinaire, lié à l'élaboration de la construction scientifique ; un sens non spécifique, celui que le nom ou l'adjectif présente dans les usages non spécialisés de la langue, décrit par exemple dans les dictionnaires usuels tels que le *Larousse* ou le *Trésor de la langue française*. Les auteurs mettent en évidence une contribution importante du LST, en particulier de certaines classes sémantiques de noms, à la formation de ces termes de type Nom-Adj. Leur analyse porte ensuite sur l'interaction en contexte du LST et des candidats-termes, poursuivant là des travaux menés antérieurement [JACQ 13]. Les auteurs confirment la tendance des candidats-termes à s'avérer effectivement des termes lorsqu'ils sont en situation de dépendance syntaxique avec une unité du LST, mais indiquent la nécessité de raffiner les observations selon les classes sémantiques de LST.

Les classes sémantiques et les relations syntaxiques en contexte sont au cœur de l'étude de Yan, présentée au chapitre 5, et qui ouvre la seconde partie de l'ouvrage. Cette seconde partie rassemble des études menées dans une perspective didactique, pour l'enseignement de l'écriture scientifique en français à des étudiants allophones, dans le cadre du français sur objectifs universitaires (chapitres 5, 6 et 8) ou pour l'enseignement de la langue scientifique aux élèves d'école primaire (6-11 ans), dans leur langue maternelle (chapitre 7).

Le chapitre 5 s'intéresse en particulier aux verbes du LST. Le problème sous-jacent traité par Yan est celui d'une organisation et d'une description des verbes français du LST en vue d'une meilleure appropriation par les scripteurs allophones devant produire des écrits académiques et/ou scientifiques. Cette description s'intègre dans la ressource exposée au chapitre 1 par Hatier. Elle permet de cerner finement les emplois des verbes présents dans l'écrit scientifique et de les associer à des classes sémantiques. Yan s'appuie pour cela sur les études des verbes du français réalisées par Dubois [DUB 97] et adopte comme modèle de description la *Corpus Pattern Analysis* proposée par Hanks [HAN 13]. Ce modèle, à la base du *Pattern Dictionary of English Verbs*, est adapté par Yan pour rendre compte des constructions présentées par les diverses acceptions des verbes du LST en termes de structures syntaxiques et d'associations lexicales régulières avec d'autres unités. Dans la tradition des classes d'objets [GRO 94, LEP 98], ce sont les propriétés syntaxiques et sémantiques des verbes qui constituent leurs descripteurs, saisies et organisées en fonction de leurs réalisations dans le corpus transdisciplinaire constitué pour toutes ces études (chapitres 1, 3, 5, 6, 8). La ressource résultante peut être exploitée par les enseignants de FLE (français langue étrangère) ou de FOU (français sur objectifs universitaires) pour construire des exercices de compréhension aussi bien que de production, dont plusieurs exemples sont donnés en fin de chapitre.

Dans la même veine, le chapitre 6 s'intéresse aux adverbes, compagnons fréquents des verbes, mais aussi des phrases ou des propositions. Les adverbes et syntagmes adverbiaux sont, dans l'écrit scientifique, vecteurs aussi bien de modalisation, donc concernent la façon dont l'auteur scientifique se situe vis-à-vis de ce qu'il écrit, que de modification du contenu propositionnel, qu'ils précisent ou au contraire tirent vers la généralité ou l'approximation. Ils représentent alors de précieux auxiliaires pour l'écriture en même temps que de redoutables écueils pour le scripteur allophone. Tran précise d'abord dans ce chapitre les raisons pour lesquelles les descriptions et typologies déjà proposées pour les adverbes et syntagmes adverbiaux, aussi pertinentes soient-elles, ne sont pas adaptées aux scripteurs allophones. Ceux-ci ne perçoivent pas aisément les nuances sémantiques et rhétoriques (telles que celles qui sont dévoilées dans le chapitre 2 par Grossmann) qui guident le choix de tel ou tel adverbe. Leur besoin est d'ordre onomasiologique : étant donné tel contenu à exprimer, quelles formes sont disponibles ? L'auteure élabore dans ce chapitre une typologie apte à satisfaire ce besoin, puisque fondée sur les caractéristiques fonctionnelles des adverbiaux. Dans la continuité de ses travaux précédents, elle place le projecteur sur certaines classes pour en préciser les propriétés. Des exemples concrets d'exercices proposés aux étudiants, en compréhension autant qu'en production d'écrits, illustrent le caractère opératoire de cette typologie.

Le chapitre 7 quitte le contexte de l'enseignement supérieur et de l'apprentissage du français comme langue étrangère en complétant les études de corpus sur le LST par une analyse de manuels scolaires du primaire. Il s'agit de mettre en œuvre les techniques d'extraction du LST, similaires à celles qui sont décrites dans le chapitre 1, pour dresser la liste des unités du LST employées dans les manuels destinés aux élèves de 8 à 10 ans en contexte scolaire québécois, et de construire ainsi une ressource pour les enseignants du primaire qui doivent définir le vocabulaire sur lequel travailler prioritairement avec les élèves en enseignement-apprentissage de la langue maternelle. Les unités du LST apparaissent par principe « rentables » pour un enseignant dans la mesure où elles vont se retrouver avec le même sens dans diverses disciplines. Pour les inventorier, Tremblay, Saidane et Drouin ont sélectionné et analysé treize manuels couvrant quatre disciplines (au sens large). Le chapitre décrit le détail de la procédure d'extraction à l'issue de laquelle 67 noms, 50 verbes et 15 adverbes émergent. La comparaison avec de semblables listes construites à partir de l'écrit scientifique « adulte » met en évidence une intersection partielle et montre l'intérêt d'un enseignement-apprentissage ciblé de ce vocabulaire, socle de l'écriture scientifique à tous les niveaux de connaissance.

Dernier de l'ouvrage, le chapitre 8, sous la plume de Cavalla, donne des pistes didactiques très précises pour l'enseignement du LST en contexte universitaire. Les neurosciences ont élargi notre compréhension de la façon dont le cerveau gère les connaissances lexicales. Le modèle le plus apte à saisir l'organisation des connaissances lexicales n'est pas celui du rangement dans des cases ou des tiroirs, que l'on ouvrirait

selon son besoin d'expression ou de compréhension, mais plutôt celui du réseau et des connexions. La compréhension et la mémorisation du lexique passeraient par l'établissement et/ou le renforcement de chemins reliant des items lexicaux. Une difficulté en langue étrangère est d'éviter le recours à la traduction, inopérant face à des séquences non compositionnelles et/ou présentant un certain figement, telles *(se) poser une question*, que certains apprenants ont envie d'énoncer sous la forme *demander une question*, par calque sur l'anglais *to ask a question.* Le parti résolument pris par les propositions didactiques formulées est de travailler les unités cibles de l'apprentissage dans leur contexte, donc en corpus, démarche qui s'inscrit pleinement dans le champ du *learning with corpora* [CHA 07, BOU 14]. Un outil tel que Scientext, rassemblant un corpus d'écrits scientifiques et une interface d'interrogation, originellement destinés à l'étude linguistique, se laisse aisément détourner pour la mise au point de séquences d'apprentissage du lexique. Cavalla, comme Tran dans le chapitre 6, insiste sur l'importance d'une double approche du lexique : sémasiologique, qui permet d'accéder au sens des formes, et onomasiologique, qui part du besoin d'expression et de la fonction pour proposer les formes adéquates. L'écriture scientifique, dont la dimension rhétorique et argumentative est cruciale, peut ainsi être plus facilement maitrisée.

De l'inventaire à l'enseignement, de l'extraction à l'analyse, d'une perspective lexicographique et lexicologique à une perspective didactique, l'ensemble des chapitres, croisant les approches et les préoccupations des chercheur-e-s, montrent le poids du lexique scientifique transdisciplinaire dans l'écriture scientifique, en particulier dans les sciences dites humaines et sociales.