Annotation du lexique scientifique transdisciplinaire dans des écrits scientifiques en Sciences Humaines et Sociales

Évelyne Jacquey (1), Sylvain Hatier (2),
Agnès Tutin (2), Laurence Kister (1)(3)

(1) ATILF – Analyse et Traitement Informatisé de la Langue Française, Nancy
(2) LIDILEM – Linguistique et Didactique des Langues Étrangères et Maternelles, Grenoble
(3) Université de Lorraine, Nancy, evelyne.Jacquey@atilf.fr, sylvain.hatier@univ-grenoble-alpes.fr, laurence.kister@univ-lorraine.fr, agnes.tutin@univ-grenoble-alpes.fr

Résumé : Dans cet article, nous présentons une expérience d'annotation manuelle en corpus d'occurrences du lexique scientifique transdisciplinaire en Sciences Humaines et Sociales (SHS). Le choix du domaine des SHS est intéressant dans la mesure où il soulève des difficultés de délimitation lexicale entre le lexique disciplinaire des disciplines des SHS dont relèvent les articles du corpus, le lexique scientifique transdisciplinaire des SHS et les lexiques de langue générale. Ces frontières floues entre les différents lexiques qui sont présents dans les articles scientifiques annotés se révèlent précisément lorsqu'on tente d'annoter et de désambiguïser les occurrences du lexique scientifique transdisciplinaire. Malgré cette difficulté, la ressource annotée pourra être particulièrement utile pour illustrer en corpus le fonctionnement du Lexique Scientifique Transdisciplinaire.

Mots-clés : annotation de corpus, lexique scientifique transdisciplinaire, terminologie

1 Introduction

Dans le domaine des sciences humaines et sociales (SHS par la suite), l'identification et la modélisation du lexique scientifique transdisciplinaire (LST par la suite) posent des difficultés particulières pour plusieurs raisons.

La première est que, dans le cas des SHS, l'intersection entre le LST et les langues de spécialité, c'est-à-dire le vocabulaire spécifique de chaque discipline, qu'on pourrait assimiler au vocabulaire terminologique dans les domaines techniques, est diversifiée et étendue. Cela conduit à de nombreux cas d'ambiguïté entre acception disciplinaire et acception transdisciplinaire. Il en est ainsi du nom *sujet*, qui relève du LST dans nombre de disciplines des SHS, mais qui peut être ambigu avec le terme correspondant

en linguistique, lorsqu'un article fait référence à la fonction grammaticale du même nom. Ce type d'ambiguïté apparaît aussi de manière régulière avec le nom *objet* ou encore avec le nom *rôle* quand un article de linguistique traite des rôles thématiques tels que les développe Fillmore.

Cette difficulté apparaît de manière plus complexe encore lorsque les entrées du LST sont en intersection avec des termes de disciplines différentes pouvant correspondre à des concepts partagés entre disciplines, des concepts proches, ou des concepts différents (Jacquey et al. 2013; 2014). On en trouve un exemple avec le nom *valeur*, dont la signification est jugée disciplinaire en psychologie dans le terme complexe *valeurs d'égalité*, mais qui renvoie à un concept différent d'une autre acception disciplinaire en linguistique, celle que l'on trouve dans des termes complexes comme *valeur de vérité* ou *valeur sémantique*.

La seconde raison du choix des SHS est que les noms du LST ont une tendance forte à entrer dans la composition de termes complexes de forme nominale étendue (Jacquey et al. 2018), c'est-à-dire composés d'un nom modifié par un adjectif ou d'un nom suivi d'un complément du nom, par exemple dans *analyse syntaxique*. Il en est ainsi du nom *corpus* dans des expressions comme *corpus de référence*, *corpus de test*, *corpus représentatif*, qui sont caractéristiques de la linguistique de corpus ou du traitement automatique des langues. Le même type de difficulté apparaît avec le nom *dépendance* dans des expressions comme *analyse en dépendance* et *arbre de dépendance* qui sont caractéristiques d'un courant théorique de la linguistique s'intéressant à la modélisation syntaxique des langues.

La troisième source d'ambiguïté vient du LST lui-même. Le LST contient, en effet, beaucoup d'entrées qui ont plusieurs acceptions transdisciplinaires entre lesquelles il est parfois difficile d'arbitrer, notamment dans le cas d'adverbes (par exemple, *donc* ou *aussi*), d'adjectifs (par exemple, *étroit* ou *proche*) et de verbes (par exemple, *considérer*, *désigner*, *intégrer*, *situer*), ainsi que, dans une moindre mesure, de noms (par exemple, *connaissance*, *indice*, *résultat*).

Une dernière raison du choix des SHS est que le LST apparaît de manière beaucoup plus ciblée et bien moins ambiguë dans les sciences dites exactes. Dans un corpus comportant 530 résumés d'articles de chimie, les cas d'ambiguïté entre acception transdisciplinaire et acception disciplinaire semblent moins nombreux. Dans ce corpus, les noms *agent*, *phase* ou *réduction* ne sont jamais employés avec leur acception transdisciplinaire. De plus, la diversité des entrées transdisciplinaires présentes est moindre. On rencontre essentiellement des adjectifs comme *relatif* ou *voisin*, peu d'adverbes simples comme *aussi*, des expressions transdisciplinaires comme *en l'absence de*, *en présence de*, des noms comme *analyse*, *comportement*, *résultat*, et quelques verbes ambigus au sein du LST comme *démontrer*, *déterminer*, *développer*, *expliquer*, *montrer*. Il apparaît donc intéressant de comparer le degré d'ambiguïté des textes de « sciences dures » avec celui que l'on observe en SHS.

Dans le cadre du projet TermITH, le LIDILEM a mis en œuvre les travaux permettant l'identification et la modélisation du LST dans le domaine des SHS. Cependant, il est apparu pertinent d'observer et d'illustrer le fonctionnement de ces ressources lexicales

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

en contexte. La présente étude s'intéresse ainsi à l'annotation du LST en corpus. Pour ce faire, les occurrences de ce lexique ont été identifiées automatiquement dans trois articles relevant de trois disciplines de SHS : linguistique, psychologie et sciences de l'éducation. Cette expérience a aussi été une occasion de tester le pouvoir descriptif et discriminant des descriptions lexico-sémantiques disponibles dans le LST.

Dans la suite de cet article, la section (2) est consacrée à la présentation du LST (mode de constitution et organisation sémantique) ainsi qu'à l'identification automatique de ses occurrences dans les articles annotés. La section (3) détaille le processus d'annotation (guide et méthodologie, environnement d'annotation) ainsi que les résultats obtenus. La section (4) conclut cette étude et ouvre plusieurs perspectives.

2 Lexique scientifique transdisciplinaire

À la suite de (Tutin, 2006), nous définissons le LST comme le lexique renvoyant à l'argumentation et au discours sur les objets et procédures de l'activité scientifique (Hatier 2016 ; Hatier 2018 ; Tutin et Jacques 2018 ; Drouin 2007 ; Paquot 2010 ; Pecman 2004) [1]. Ce lexique intègre des unités lexicales comme *hypothèse* ou *analyser*, mais aussi des expressions comme *analyse de discours* ou *point de vue*. Bien que mobilisé dans un usage spécialisé de la langue, le LST n'est pas un lexique de spécialité, mais plutôt un lexique propre à un genre textuel. Il est ainsi transdisciplinaire, contrairement à la terminologie, et surreprésenté dans l'écrit scientifique, ce qui le distingue de la langue générale. Afin d'illustrer ce lexique aux frontières floues, nous présentons dans la figure (1) un extrait d'un article d'économie où nous avons mis en relief les différents lexiques présents dans les écrits scientifiques (Hatier 2016 ; 2018).

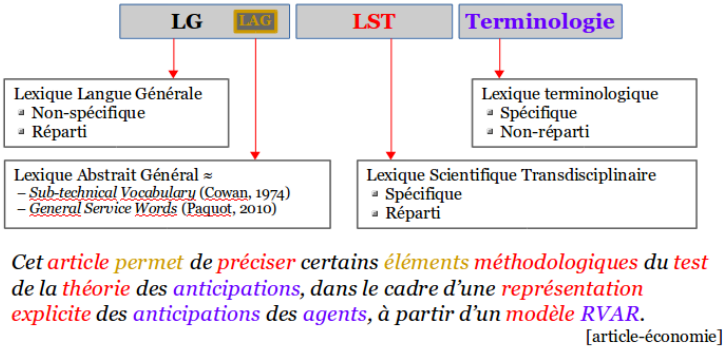


FIGURE 1: Lexiques présents dans l'écrit scientifique (Hatier 2016)

L'écrit scientifique mobilise ainsi quatre principaux lexiques : celui de la langue générale (LG), le lexique terminologique lié à chaque discipline, un lexique abstrait général (LAG) et le LST. Le LAG se distingue du LST par le fait qu'il ne présente pas de caractère spécifique au domaine scientifique ; il a été étudié sous le nom de *General Service Words* (Paquot, 2010) ou *Sub-Technical Vocabulary* (Cowan, 1974). Présentant des rôles comparables dans l'écrit scientifique, le LAG et le LST sont regroupés

1. Pour un état de l'art sur la question du LST, nous renvoyons à Tutin & Jacques (2018).

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

sous l'étiquette LST. Cette fusion se justifie également par la difficulté à distinguer deux lexiques qui partagent des propriétés lexicométriques essentielles : la transdisciplinarité et la surreprésentation dans le genre académique (en comparaison avec les genres littéraire ou journalistique).

2.1 Constitution du lexique

La constitution du lexique se fait selon une approche dite de *linguistique outillée* (Habert, 2004), composée de deux étapes principales (pour une description plus détaillée, voir Hatier, 2016). Lors d'une première étape, un ensemble de traitements automatiques permet de faire émerger d'un corpus d'écrits scientifiques des phénomènes linguistiques (occurrences de mots, cooccurrences syntaxiques). Ces éléments sont ensuite analysés manuellement afin de valider les entrées du LST. L'objectif est de produire une description fine du LST en extension, par l'inventaire de ses éléments, et en intension par la description des propriétés de ces éléments.

Les premières propriétés utilisées pour l'identification du LST sont donc statistiques. Le LST étant transdisciplinaire et spécifique au domaine scientifique, ses occurrences dans un corpus d'écrits scientifiques doivent être largement réparties dans les disciplines et statistiquement surreprésentées par rapport à d'autres genres d'écrits. Nous utilisons ainsi un corpus de contraste (représentant le genre littéraire, le genre journalistique et l'oral) pour faire émerger le LST de notre corpus d'analyse (composé de 500 articles en SHS).

Le schéma ci-dessous illustre ce processus d'identification du LST.

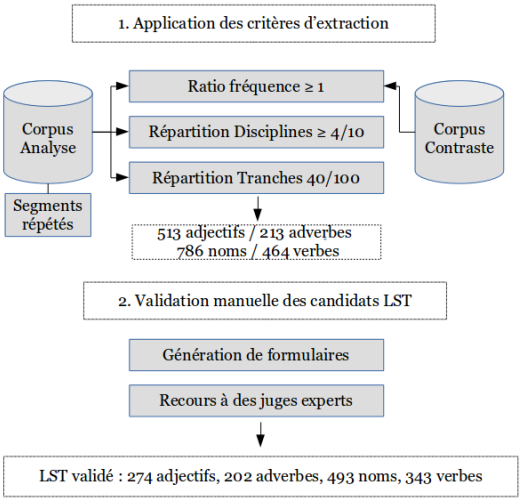


Figure 2: Processus de sélection des éléments du LST (Hatier 2016)

Les éléments vérifiant les critères statistiques sont ensuite validés manuellement par des juges experts du domaine de l'écrit scientifique. Les éléments du LST résultant de cette première étape sont ensuite analysés sémantiquement. Nous procédons, en

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

premier lieu, à l'inventaire des acceptions transdisciplinaires pour chaque lemme. Ces acceptions sont ensuite organisées dans une classification sémantique à deux niveaux en utilisant une fois encore les propriétés linguistiques observées en corpus. Nous tirons ainsi parti des fonctionnalités du *Lexicoscope* (Kraif & Diwersy, 2012 ; Kraif, 2016) à l'aide duquel nous identifions les éléments du LST partageant des propriétés distributionnelles afin de les regrouper dans des sous-classes et classes sémantiques.

L'ensemble des traitements est résumé dans la figure (3) ci-dessous.

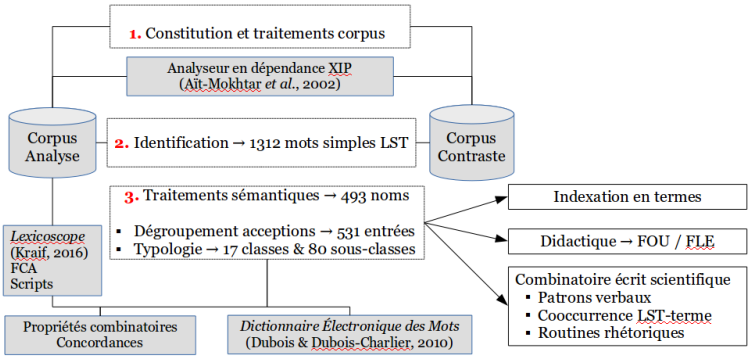


FIGURE 3: Traitement sémantique du LST

Le typage sémantique du LST répond à plusieurs besoins précédemment identifiés dans les travaux sur ce lexique. L'ajout d'information sémantique apparaît ainsi utile non seulement pour l'amélioration de l'indexation en termes, mais également pour l'appropriation de ce lexique par les scripteurs, en permettant un accès onomasiologique utile pour les applications didactiques. De plus, plusieurs études sur la phraséologie dans l'écrit scientifique ont mis en avant l'intérêt du LST, notamment pour l'étude de structures sémantico-syntaxiques.

La première tâche de dégroupement des lemmes du LST en acceptions transdisciplinaires se fait par observation des concordances en prenant appui notamment sur le *Dictionnaire Électronique des Mots* (Dubois & Dubois-Charlier, 2010). Pour chaque entrée est renseignée, outre le lemme et la catégorie, une courte glose correspondant à l'acception transdisciplinaire relevée en corpus. La deuxième étape, consistant en la création d'une classification sémantique, utilise les propriétés distributionnelles des entrées du LST pour décrire ce lexique en termes de classes et sous-classes sémantiques, en fournissant trois types d'informations : une liste des membres, une définition et un test linguistique d'appartenance.

Ainsi, si nous prenons l'exemple des noms *ouvrage* et *article,* nous avons pu observer en corpus de nombreuses cooccurrences avec les verbes *lire* et *publier.* Lors de la création de la sous-classe {document} à laquelle ils appartiennent, nous définissons un test du type '*Lire, publier un N'* permettant de valider ou non l'appartenance à la sous-classe d'un élément du LST.

Nous aboutissons alors à une ressource du LST dont les entrées sont des acceptions désambiguïsées, intégrées dans une classe et une sous-classe sémantique.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Le tableau (1) ci-dessous présente un extrait de la classification sémantique du LST.

Tableau 1 : Description d'une classe sémantique du LST avec 4 de ses sous-classes

Classe (17)	Sous-Classe (80)	Noms – Acceptions (531)
communication_support : 'acte de communication et de transfert des idées (selon le moyen de transmission)' *Le N est consacré à*	Document 'document écrit' *Lire, publier un N*	*article, document, documentation, essai_1, littérature, note, ouvrage, publication, synthèse_1 texte, thèse_1, volume_1*
	Evénement 'événement scientifique' *un N a lieu / Pendant un N*	*colloque, conférence*
	Graphique 'représentation graphique éclairant le texte' *Le N {numéral} montre / cf. N, sur N, dans N*	*figure_1, illustration_1, image_1, motif_1, tableau*
	Section 'sous-partie d'un document écrit' *Dans le N de l'ouvrage / cf. N*	*annexe, bibliographie, chapitre, conclusion_1, développement_1, extrait, introduction_1, section_1*

2.2 Projection et identification des occurrences

La procédure d'identification des occurrences a été réalisée avec Nooj (Silberztein, 2015). Pour ce faire, le LST a été transformé en un dictionnaire de formes fléchies compatible avec l'environnement Nooj. Une fois l'article au format XML-TEI-P5 et le dictionnaire chargés dans l'environnement Nooj, une analyse linguistique est lancée afin d'identifier les occurrences du LST en fonction d'un certain nombre de patrons qui sont référencés dans des objets que Nooj désigne comme des grammaires. Celles-ci contiennent des grammaires, comme celle que l'on observe ci-dessous pour les adverbes polylexicaux (ex : *pour ainsi dire, en d'autres termes).*



Figure 4: Exemple de règle de grammaire Nooj

À l'issue de cette première étape, les occurrences reconnues en fonction des règles de la grammaire Nooj sont automatiquement balisées dans l'article au format XML et l'ensemble est sauvegardé pour être utilisé lors de l'étape suivante.

3 Annotation

L'annotation manuelle est destinée à examiner et à analyser les occurrences du LST telles qu'elles ont été identifiées avec Nooj. Comme cela a été rappelé dans l'introduction, l'identification automatique qui sélectionne des successions de formes fléchies sur la base de règles de grammaire conduit inévitablement au balisage d'occurrences qui ne relèvent pas du LST. Par ailleurs, lorsque les occurrences balisées relèvent effectivement du LST, il est souvent nécessaire d'arbitrer entre les différentes acceptions transdisciplinaires possibles.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

L'annotation manuelle mise en œuvre est réalisée dans l'environnement Oxygen en mode « Author ». Chaque document annoté est associé à un schéma et une feuille de style. Le schéma est défini à l'aide d'une DTD et la feuille de style utilise le langage CSS. Cette dernière permet d'afficher de manière facile à discriminer visuellement les occurrences du LST candidates (police en vert), les cas d'ambiguïté entre plusieurs acceptions (surlignement en jaune) ainsi que les bornes de chaque occurrence reconnue sous la forme de triangles ouvrants et fermants [2].

La figure ci-dessous représente un extrait de l'article de linguistique tel qu'il apparaît pour l'annotateur avant le démarrage de l'annotation manuelle.

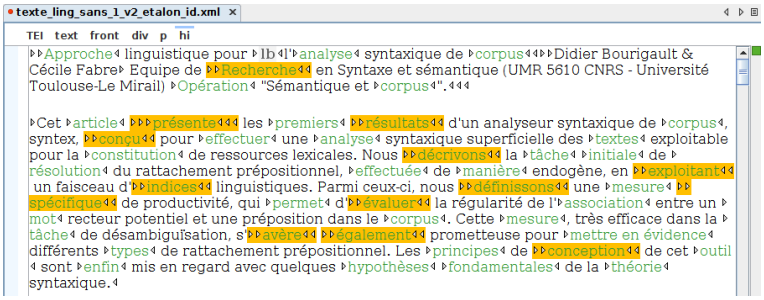


FIGURE 5: Balisage des occurrences du LST sous Oxygen avant annotation manuelle

Chaque occurrence du LST est identifiée à l'aide de balises qui ont toutes un jeu d'attributs défini par la DTD. Pour toute occurrence du LST validée lors de la phase d'annotation, les recommandations d'annotation précisent les consignes concernant les attributs « TERME », « SPE », « COMM » et « COMM*LIBRE ». Les valeurs des sept autres attributs sont directement reportées depuis le dictionnaire correspondant au LST sous Nooj ou proviennent de l'import sous Oxygen des documents à annoter (identifiant « xml :id »).

2. À noter que les triangles représentent l'ensemble des balises XML du document vu en mode « Author ». Certains triangles peuvent indiquer d'autres balises que les bornes d'une occurrence du LST, mais dans ce cas, la police de caractères n'est pas colorée en vert.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

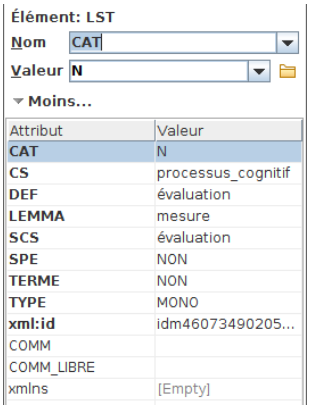


FIGURE 6: Attributs de chaque occurrence du LST

3.1 Corpus annoté

Le corpus annoté se compose de trois articles scientifiques intégraux relevant de trois disciplines différentes des SHS : linguistique, psychologie et sciences de l'éducation. Nous rappelons ici que le LST est structuré en acceptions. Autrement dit, chaque entrée du LST correspond à un sens transdisciplinaire. Ainsi, pour un certain nombre de lexèmes (forme lemmatisée associée à une catégorie grammaticale), plusieurs correspondances LST sont possibles. De ce fait, les lexèmes suivants, par exemple l'adverbe *donc* (2 acceptions), l'adjectif *proche* (3 acceptions), le nom *recherche* (2 acceptions) et le verbe *considérer* (3 acceptions) correspondent à 2 entrées LST pour *donc* et *recherche*, et à 3 entrées LST pour *proche* et *considérer*. Par conséquent, si ces 4 lexèmes sont reconnus chacun une fois dans un texte, il y aura 10 occurrences candidates du LST qui feront l'objet d'une annotation. Le tableau (2) fournit les nombres d'occurrences des entrées LST par discipline et par catégorie grammaticale, ainsi que la proportion d'occurrences LST ambiguës au sein du lexique.

TABLEAU 2 : Métriques du corpus annoté

Discipline	Tokens	Total des occurrences du LST				Dont occurrences ambiguës			
Linguistique	6 994	2 526 *(36,12 % des tokens de l'article)*				1 075 *(42,56 % du total des occurrences LST)*			
		270	192	1 349	715	93	76	354	552
		10,69 %	*7,60 %*	*53,40 %*	*28,31 %*	*3,68 %*	*3,01 %*	*14,01 %*	*21,85 %*
Psychologie	3 119	1 019 *(32,67 % des tokens de l'article)*				455 *(44,65 % du total des occurrences LST)*			
		162	92	431	334	63	44	86	262
		15,90 %	*9,03 %*	*42,30 %*	*32,78 %*	*6,18 %*	*4,32 %*	*8,44 %*	*25,71 %*
Sciences de l'éducation	4 960	1 458 *(29,40 % des tokens de l'article)*				571 *(39,16 % du total des occurrences LST)*			
		167	161	714	416	45	52	162	312
		11,45 %	*11,04 %*	*48,97 %*	*28,53 %*	*3,09 %*	*3,57 %*	*11,11 %*	*21,40 %*
Total	15 073	5 003 *(36,12 % des tokens du corpus)*				2 101 *(41,99 % du total des occurrences LST)*			
		599	445	2 494	1 465	201	172	602	1 126
		11,97 %	*8,89 %*	*49,85 %*	*29,28 %*	*4,02 %*	*3,44 %*	*12,03 %*	*22,51 %*
Catégories du LST		Adj	Adv	N	V	Adj	Adv	N	V

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Dans ce tableau, la proportion d'occurrences du LST pour chaque discipline est calculée par rapport à la taille de l'article de la discipline concernée, autrement dit par rapport au nombre de tokens de l'article. Les proportions d'occurrences par catégorie et les proportions d'occurrences ambiguës sont calculées par rapport au nombre total d'occurrences du LST. Comme on peut le constater, les proportions sont globalement similaires, la part des noms du LST se situant aux alentours de 50 %, la part des verbes autour de 30 % et la part des adjectifs et des adverbes autour de 10 %. Parmi ces occurrences, la proportion de celles qui sont ambiguës par rapport au nombre total d'occurrences LST est en moyenne d'environ 23 % pour les verbes, de 12 % pour les noms, de 7 % pour les verbes et de 4 % environ pour les adjectifs et les adverbes. Cette distribution classique montre que les noms LST sont environ deux fois plus fréquents que les verbes et deux fois moins ambigus.

3.2 Consignes d'annotation et déroulement de l'annotation

L'annotation vise la sélection des occurrences qui relèvent effectivement du LST ainsi que, dans la mesure du possible, la sélection de l'acception qui semble la plus appropriée en fonction du contexte de l'occurrence parmi les différentes acceptions transdisciplinaires possibles.

L'annotation manuelle a été réalisée par quatre annotateurs pour chacun des trois articles. Les annotateurs ont un profil d'experts dans la tâche d'annotation : ils disposent d'une formation de niveau avancé en linguistique et connaissent bien la problématique du lexique scientifique transdisciplinaire. Deux grandes étapes ont été définies dans le processus d'annotation. La première a consisté à éliminer les occurrences qui ne relevaient pas d'une acception transdisciplinaire, la seconde a consisté à la fois à sélectionner les acceptions transdisciplinaires appropriées au contexte de l'occurrence annotée et à saisir les valeurs des attributs le cas échéant.

3.2.1 Élimination des occurrences non transdisciplinaires

Plusieurs cas de figure ont conduit à l'élimination d'occurrences correspondant potentiellement à une entrée du LST. Cette élimination consiste à supprimer la balise LST.

Les entités nommées

Certaines occurrences du LST peuvent apparaître à l'intérieur d'une entité nommée. C'est par exemple le cas de l'occurrence *recherche* dans l'entité *Équipe de Recherche en Syntaxe et Sémantique*.

Les occurrences faisant partie de séquences plus étendues

— Occurrences incluses dans des déterminants complexes ou des locutions prépositionnelles ou conjonctives

Seules les occurrences d'adjectifs, d'adverbes, de noms et de verbes sont annotées, car le LST comporte seulement ces quatre catégories simples, auxquelles s'ajoutent quelques expressions polylexicales assimilables à l'une des quatre catégories simples. De ce fait, toutes les occurrences du LST apparaissant au sein de déterminants com-

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

plexes (par exemple, *certain* et *nombre* dans *un certain nombre de X*), de locutions prépositionnelles (par exemple, *fonction* dans *en fonction de*) et de locutions conjonctives (par exemple, *même* dans *de même que*) sont supprimées.

— Occurrences incluses dans une occurrence LST polylexicale

Ce cas de figure se présente avec l'adverbe *plus* lorsqu'il apparaît dans la locution adverbiale *de plus* ou encore avec le nom *ensemble* dans la locution adverbiale *dans son ensemble*. La même situation se produit avec le nom *base* dans la locution adjectivale *de base*. Dans ce type de configuration, si l'annotateur était bien face à la locution ou à l'expression polylexicale figurant dans le LST, alors il avait pour consigne d'éliminer l'occurrence simple et de conserver l'occurrence polylexicale.

— Occurrences ambiguës avec une acception disciplinaire

Ce cas de figure se présente différemment selon les disciplines. Ainsi, en linguistique, cette situation apparaît avec des noms comme *corpus*, *mot*, *texte*. Il en est de même pour les noms *minorité* ou *valeur* dans le texte de psychologie qui s'intéresse au racisme et à ses modes d'expression[3].

— Occurrences mal catégorisées

Un certain nombre d'entrées du LST sont catégoriellement ambiguës. C'est le cas d'unités comme *certain*, *différent*, *divers* qui sont ambiguës entre les catégories déterminant, adjectif et parfois nom. De même, il existe une ambiguïté bien connue entre passifs réduits de verbe, verbe au participe passé et adjectif, par exemple, avec les mots *accru* ou *approfondi*). La consigne d'annotation est de vérifier la conformité entre la catégorie grammaticale que l'annotateur attribue à l'occurrence en contexte et la catégorie grammaticale issue du LST et associée à l'occurrence par Nooj. S'il n'y a pas conformité ou si l'occurrence relève d'une autre catégorie que les quatre catégories majeures qui sont présentes dans le LST, la consigne est alors de supprimer la balise .

3.2.2 Annotation des occurrences transdisciplinaires conservées et sélection de l'acception transdisciplinaire appropriée

Dans cette seconde étape de l'annotation manuelle, il ne reste que des occurrences relevant d'acceptions transdisciplinaires, linguistiquement bien formées et catégoriellement conformes à la fois au contexte des occurrences et à la catégorie grammaticale fournie pour la ou les entrée(s) correspondante(s) dans le LST. À ce stade, deux cas de figure n'ont pas encore reçu de recommandations d'annotation : les cas d'ambiguïté entre plusieurs acceptions transdisciplinaires et les cas où le balisage *LST* est conservé avec certaines précisions qui sont encodées *via* les attributs.

En ce qui concerne les cas d'ambiguïté, l'annotateur est appelé à utiliser une approche intuitive tout en ayant la possibilité de s'appuyer sur un recueil d'exemples d'usage des différentes entrées du LST ainsi que sur une description de la sémantique des classes

3. Dans ce cas de figure, les consignes d'annotation étaient initialement d'associer la valeur « OUI » à l'attribut « SPE ». À l'issue d'une étape de test, la consigne finale a été de supprimer la balise *LST* pour les occurrences ambiguës avec une acception disciplinaire.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

et sous-classes sémantiques sous-jacentes à la structuration sémantique du LST. De plus, deux heuristiques ont été proposées. La première concerne l'ambiguïté entre une interprétation qualitative et une interprétation quantitative telle qu'elle apparaît avec les deux entrées du nom *résultat* dans le LST. En cas de doute, l'annotateur est invité à privilégier l'interprétation la plus large possible qui est l'interprétation qualitative. Un autre type d'ambiguïté récurrent dans le LST apparaît entre une interprétation supposant un agent humain et une interprétation ne le supposant pas. Le verbe *tendre*, par exemple, présente ce type d'ambiguïté. L'entrée *tendre*1* fait référence à un sens supposant un agent humain et qui est glosé par *porter vers un but* tandis que l'entrée *tendre*2*, ne supposant pas un agent exclusivement humain, est glosée dans le LST par *évoluer vers un état différent*. Dans ce cas, si l'annotateur hésite sur l'interprétation à privilégier, la consigne est de choisir préférentiellement l'interprétation supposant un agent humain.

Le dernier cas de figure concerne des occurrences pour lesquelles le balisage est maintenu, mais pour lequel des précisions doivent être fournies. Une première situation générale entre dans ce cas de figure : il s'agit de tous les cas où l'annotateur estime nécessaire de commenter sa décision. Il dispose pour cela de deux attributs. L'attribut « COMM » permet d'encoder des commentaires dont les valeurs ont été prédéfinies :

— « MAN » fait référence au fait qu'il manque une acception pour décrire le sens que perçoit l'annotateur en contexte,
— « DEF » fait référence au fait que la définition ou la glose fournie par le LST est inadaptée,
— « CS » et « SCS » ont la même utilité concernant respectivement la classe sémantique et la sous-classe sémantique,
— « AMB » permet à l'annotateur d'indiquer que l'ambiguïté qu'il perçoit pour l'occurrence traitée est très difficile à résoudre.

Un second attribut, « COMM_LIBRE », permet à l'annotateur d'indiquer des précisions qu'il juge utiles et qui n'ont pas été prédéfinies. La seconde situation dans laquelle l'annotateur est invité à utiliser les attributs pour fournir des précisions autres que des commentaires correspond au cas où une occurrence monolexicale transdisciplinaire est incluse dans un terme complexe, par exemple, le nom *analyse* dans le terme *analyse syntaxique* en linguistique. Dans ce dernier cas, le balisage *LST* est maintenu autour de l'occurrence monolexicale transdisciplinaire et la valeur « OUI » est associée à l'attribut « TERME ».

3.3 Résultats et analyse

L'ensemble des annotations produites par les quatre annotateurs ont été fusionnées afin de calculer les taux d'accord inter-annotateurs. Dans la présente expérience, nous avons utilisé le Kappa de Fleiss conformément aux recommandations que l'on trouve dans de nombreux travaux au sujet de l'évaluation des annotations manuelles, parmi d'autres (Mathet et Widlöcher 2016). Le principe de cette mesure est de corriger l'accord que l'on mesure entre les annotateurs (l'accord observé, Ao) en fonction de l'accord auquel on aurait pu s'attendre à partir de la distribution des annotations

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

réalisées (accord attendu, *Ae*). Le coefficient obtenu, dit accord inter-annotateur, est le résultat de la formule suivante.

$$\kappa = \frac{(Ao - Ae)}{1 - Ae}$$

La mesure globale d'accord inter-annotateur (encore appelé κ ou *kappa* dans la littérature) sur l'ensemble des occurrences candidates du LST est de 0,5588 en linguistique, de 0,6066 en psychologie et de 0,4341 en sciences de l'éducation pour des accords observés comparables dans les trois disciplines : 0,8009 en linguistique, 0,8297 en psychologie et 0,8057 en sciences de l'éducation. Le *kappa* en sciences de l'éducation est plus faible que dans les autres disciplines du fait que l'accord attendu y est plus élevé (0,6566 par rapport à 0,5487 et 0,5671 pour les deux autres disciplines). Ce premier résultat montre que l'annotation produite s'approche d'une qualité correcte au regard de l'interprétation habituelle des Kappas qui établit qu'un accord au-dessus de 0,80 dénote une annotation de très bonne qualité et un accord se situant entre 0,60 et 0,80 indique une annotation de qualité correcte. Selon (Mathet et Widlöcher 2016) parmi d'autres, l'intérêt de ce type de mesure est d'évaluer si l'annotation produite a de bonnes chances d'être reproductible et constitue donc une annotation dite de référence.

Bien que ce type de mesure soit très couramment employé dans les travaux traitant de l'annotation, celle-ci ne correspond pas tout à fait à l'objectif de l'expérience qui a été menée : observer le LST en contexte réel et faire émerger des difficultés d'annotation pouvant être liées au lexique lui-même. Nous nous sommes donc intéressés à des mesures plus intuitives et plus parlantes relativement à ces objectifs. La première mesure consiste à évaluer la proportion d'occurrences pour lesquelles 3 annotateurs sur 4 s'accordent pour conserver ou pour rejeter une occurrence candidate. La seconde mesure n'en est pas une à proprement parler. Il s'agit plutôt d'un relevé. Cette seconde piste d'analyse consiste à examiner plus en détail les occurrences candidates pour lesquelles aucune décision ne peut être prise parce que les choix des annotateurs sont répartis à parts égales entre conservation et rejet de l'annotation LST.

3.3.1 Observation des choix des annotateurs

Le tableau (3) ci-dessous synthétise les proportions et les effectifs des choix des annotateurs sur l'ensemble du corpus. Nous considérons, dans le cadre de cette mesure intuitive, qu'un choix est retenu à partir du moment où 3 annotateurs sur 4 ont, soit conservé l'annotation LST pour l'occurrence examinée (colonne « LST »), soit l'ont rejetée (colonne « Non-LST »). La troisième colonne reproduit le décompte (effectif et proportion) des cas où aucune conclusion n'est possible parce que 2 annotateurs sur 4 ont choisi de conserver l'annotation LST et les 2 autres ont choisi de la rejeter (colonne « Indécidable »). Nous rappelons enfin que toutes les proportions indiquées sont calculées par rapport au nombre total d'occurrences LST candidates pour chaque discipline.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

TABLEAU 3 :Distribution des choix des annotateurs entre LST et non-LST pour les trois disciplines

Modalités	Basaa	Beti-fang (ewondo)	Fufulde	Swahili
Mots traduits	73 %	25 %	0 %	83 %
Unité lexicales simple (ULs)	21,91 %	84 %	0 %	61 %

Les décomptes du tableau (3) fournissent une représentation plus synthétique et plus parlante des grandes tendances que l'on peut observer sur les choix des annotateurs concernant les 5 003 occurrences du LST candidates qui ont été examinées manuellement. Indépendamment du critère d'ambiguïté des occurrences annotées, on observe tout d'abord que les annotateurs réalisent un choix dans la très grande majorité des cas : 4 791 sont effectivement arbitrées, soit 95,76 % des cas toutes disciplines confondues. Parmi les choix partagés par au moins 3 annotateurs sur 4, le choix de conserver l'annotation est nettement dominant dans les trois disciplines : 65,44 % des occurrences candidates en linguistique, 68,79 % en psychologie et 78,33 % en sciences de l'éducation. Le choix de rejeter l'annotation LST est plus important en linguistique (30,48 %) qu'il ne l'est en psychologie (26,20 %) et enfin en sciences de l'éducation (17,70 %). Cette répartition entre LST et Non-LST se retrouve aussi au sein des occurrences LST ambiguës. Autrement dit, l'ambiguïté d'une acception LST ne semble pas particulièrement affecter les choix des annotateurs. On ne peut cependant rien conclure à ce stade concernant les disciplines parce qu'elles ne sont représentées que par un seul texte dans cette expérience. Les différences constatées entre disciplines peuvent donc être le fait des textes eux-mêmes et non des disciplines.

Les cas indécidables représentent en moyenne 4,5 % des occurrences candidates. Ces cas sont l'objet de la section qui suit.

3.3.2 Examen des cas de désaccord

Les 212 cas d'annotations non arbitrées par les annotateurs se répartissent en 168 occurrences d'entrées du LST ambiguës et 44 occurrences d'entrées du LST non ambiguës. Parmi les occurrences du LST ambiguës, on trouve une majorité de verbes (118 occurrences), une vingtaine d'occurrences de noms et d'adjectifs et 7 occurrences d'adverbes.

Le premier type de difficulté rencontré concerne une difficulté liée à la délimitation du LST. Il arrive que des unités aient été retenues alors qu'étant intégrées dans des mots grammaticaux ou des expressions polylexicales ne relevant pas du LST, elles auraient dû être écartées. C'est le cas de la conjonction *dès lors que* dans l'exemple suivant pour laquelle certains annotateurs ont retenu à tort l'adverbe *dès lors*.

(1) Autrement dit, aucun impératif lié à un gain économique en termes de diminution de coût du travail qualifié du formateur n'est en jeu, même si d'une part le temps passé doit être régulé et si, d'autre part, les observations menées avec une deuxième génération d'étudiants montrent que ce temps peut être moins important en durée dès lors que les pratiques collaboratives des étudiants entre eux s'intensifient. (article, sciences de l'éducation)

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Ce problème est assez fréquent pour les locutions conjonctives et prépositionnelles, pour lesquelles une vigilance particulière s'impose.

Le deuxième type de problème rencontré par les annotateurs est celui du choix entre des acceptions ambiguës à l'intérieur du LST. L'annotateur peut éprouver des difficultés à opérer certaines distinctions en contexte. Certaines ambiguïtés apparaissent assez faciles à traiter, car les acceptions sont assez éloignées. Tel est le cas de *figure* dans l'exemple suivant (qui a deux acceptions dans le LST : 'schéma graphique' ou 'représentation mentale') :

(2) 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 syntaxique (figure 1). (article Linguistique)

Dans d'autres cas, les acceptions apparaissent plus proches et plus difficiles à démêler. Il s'agit d'adverbes à fonction discursive comme *aussi*, *donc*, d'adjectifs comme *grand*, des noms comme *cas*, *conception*, *développement*, *mesure* et de nombreux verbes comme *considérer*, *constituer*, *développer*, *démontrer*, *évaluer*, *exprimer*, *indiquer*, *signaler* pour lesquels les annotateurs semblent avoir rencontré de grandes difficultés à arbitrer entre les acceptions possibles dans l'ensemble des disciplines de l'expérience. On peut ainsi relever pour les verbes que la distinction entre les acceptions à sujet humain ou non humain, assez systématique dans le traitement du LST, ne semble pas avoir été suffisamment identifiée par les annotateurs, faute de définition claire et d'exemples suffisants. Un exemple comme le suivant n'a ainsi pas rencontré un accord satisfaisant entre les annotateurs.

(3) Plus spécifiquement, les résultats que nous avons exposés démontrent que le critère de l'association lexicale, exprimée en termes de productivité, fournit un outil puissant pour déterminer les relations de dépendance entre mots (article Linguistique)

Le verbe *démontrer* comporte deux acceptions dans le LST : *démontrer 1* (sujet humain, classe sémantique : analyse*info, sous-classe : #démontration) et *démontrer 2*, à sujet non humain, qui relève de la classe sémantique 'processus*non*humain' (avec la sous-classe 'révélation'). Dans cet exemple, le verbe relève plutôt de la deuxième acception, car le sujet non humain n'effectue pas une analyse de l'information, mais exhibe des phénomènes. Cependant, la désambiguïsation peut s'avérer complexe avec certains exemples, car il arrive par métonymie que certains sujets non humains puissent être des sujets de l'acception 1, comme dans le cas suivant :

(4) Plusieurs études ont démontré également que les connaissances métalinguistiques en langue maternelle sont étroitement liées à la réussite de l'apprentissage de la langue étrangère. (source : [lin-art-634] – développement)

Comme nous l'indiquons plus bas, l'ajout d'informations simples sur les propriétés syntaxiques et l'affichage des collocations privilégiées des verbes, que nous souhaitons intégrer dans une version ultérieure, devraient faciliter la tâche d'annotation.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Enfin, le dernier cas d'ambiguïté touche au flou des frontières entre acception disciplinaire et acception transdisciplinaire, dans certains cas et selon les disciplines. Des lexèmes comme *corpus* ou *ambigu* relèvent-ils du lexique transdisciplinaire ou relèvent-ils du lexique disciplinaire de la linguistique ?

(5) 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). (article Linguistique)

(6) Or nous voulons que l'analyseur prenne une décision (et une seule) dans ces cas ambigus, pour identifier la structure syntaxique du syntagme et enrichir ainsi le réseau de dépendance. (article Linguistique)

Dans ces exemples, les lexèmes *corpus* et *ambigu* sont très proches de l'acception transdisciplinaire, mais ils se spécialisent ici en quelque sorte. La notion de corpus renvoie non seulement à un ensemble de textes, mais à un ensemble de productions qui présentent une forme de représentativité du langage. Le mot *ambigu* renvoie ici à une interprétation multiple associée spécifiquement à des signes linguistiques.

4 Conclusion et perspectives

Cette expérience d'annotation montre la difficulté à circonscrire la notion de lexique transdisciplinaire. Cette difficulté est illustrée par le taux d'accord moyen observé dans les résultats de l'annotation, bien que les annotateurs aient une expérience confirmée de la notion de lexique transdisciplinaire, des données lexicales elles-mêmes et de ce type d'annotation.

Une piste d'amélioration est tout d'abord une meilleure présentation de la ressource lexicale, avec des exemples pertinents. Par ailleurs, la description des classes et sous-classes sémantiques ainsi que celle des gloses ne sont pas encore suffisamment détaillées pour qu'on puisse arbitrer facilement entre des acceptions concurrentes. Autrement dit, les descriptions actuelles ne permettent pas aux annotateurs de se construire une image mentale suffisamment synthétique de chaque classe ainsi que des distinctions entre classes. Par ailleurs, étant donné le type d'annotation demandé, l'annotateur a une tendance naturelle à s'appuyer sur la partie définition ou glose des entrées dans le LST. Cependant, cette partie a été développée dans un second temps par rapport à la structuration en classes, et elle est moins avancée.

Deux cas d'ambiguïté semblent particulièrement difficiles à arbitrer. Le premier cas concerne des acceptions transdisciplinaires qui ont pu sembler trop proches aux annotateurs ou dont le contexte d'occurrence ne fournissait pas suffisamment d'informations explicites pour prendre une décision satisfaisante. Nous pensons qu'une amélioration des ressources devrait permettre de lever la plupart des ambiguïtés. Le second cas vient du fait qu'en SHS, la frontière entre lexique disciplinaire et lexique transdisciplinaire est parfois floue. Dans ce dernier cas de figure, on trouve notamment les cas d'ambiguïté entre terminologie et lexique transdisciplinaire avec des noms comme *égalité*, *sujet* ou encore *valeur*. Ce cas de figure est aussi illustré par un certain nombre d'occurrences qui sont aussi des têtes nominales de termes complexes

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*,
Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

comme *analyse*$_{LST}$ syntaxique*, *catégorie*$_{LST}$ syntaxique*, *valeur*$_{LST}$ sémantique*. La question qui se pose alors est de savoir si la tête nominale a bien son acception transdisciplinaire dans ce contexte.

À l'issue de cette expérience, plusieurs améliorations peuvent être envisagées concernant le LST,en ce qui concerne la ressource, mais aussi l'annotation. Il serait tout d'abord souhaitable d'enrichir la définition des mots du LST (en particulier, les plus ambigus) ainsi que les sous-classes et classes sémantiques, afin de faciliter la désambiguïsation. De plus, il serait pertinent d'accompagner les verbes de quelques informations syntaxiques simples, qui sont des informations indispensables pour l'analyse sémantique, et seraient par ailleurs utiles pour les applications didactiques du lexique (par exemple, qqun *attribue* qqch à qqch). En outre, les collocations du LST associées aux acceptions devraient également faciliter la désambiguïsation (par exemple, *sens opposé* pour 'sens 1' de direction, ou *saisir le sens* pour 'sens 2' de signification). Enfin, il est souhaitable que la ressource lexicale soit facilement accessible en ligne pour l'annotateur, dans un environnement ergonomique.

Concernant la méthodologie d'annotation, un travail de clarification du guide d'annotation doit être entrepris, en précisant le traitement des cas potentiellement ambigus. Une réflexion sur l'arbitrage des cas complexes doit aussi être menée. Cela permettrait notamment d'étendre le corpus annoté en LST afin de proposer une ressource utile pour les perspectives applicatives, pour l'enseignement de l'écrit scientifique ou pour le traitement automatique des langues.

Bibliographie

Cowan J. R. (1974). "Lexical and syntactic research for the design of EFL reading materials", *TESOL Quarterly*, n° 8(4 p. 389-399), ISSN : 0039-8322.

Drouin, P. (2007). « Identification automatique du lexique scientifique transdisciplinaire », *Revue française de linguistique appliquée*, n° XII(2), p. 45-64, ISSN : 1386-1204, e-ISSN : 1875-368X.

Dubois, J. ; Dubois-Charlier, F. (2010). « La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration », *Langages* (3), p. 31-56, ISSN : 0458-726X, e-ISSN : 1958-9549.

Habert, B. (2004). « Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs », *Revue française de linguistique appliquée*, n° IX(1), p. 5-24, ISSN : 1386-1204, e-ISSN : 1875-368X.

Hatier, S. (2016). « Identification et analyse linguistique du lexique scientifique transdisciplinaire. Approche outillée sur un corpus d'articles de recherche en SHS », *thèse de doctorat*, Université Grenoble Alpes, Grenoble.

Hatier, S. (2018). « Identification et analyse linguistique des noms du LST », *in* Jacques, M. P. ; & Tutin, A. (dir.). *Lexique transversal et formules discursives des sciences humaines*, ISTE Editions, Londres, p. 29-40, ISBN : 9781784054854 (papier), ISBN : 9781784064853 (ebook).

Jacquey, E. ; Tutin, A. ; Kister, L. ; Jacques, M.-P. ; Hatier, S. ; Ollinger, S. (2013). « Filtrage terminologique par le lexique transdisciplinaire scientifique : une expérimentation en sciences humaines », *TIA 2013*, Villetaneuse, France, p. 121-128.

Jacquey, E. ; Kister, L. ; Humbert, J.-M. ; Tutin, A. (2014). « Terminologie et phraséologie transdisciplinaire dans les articles scientifiques en SHS », *ACFAS 2014*, Montréal, Canada.

Jacquey, E. ; Kister, L. ; Marcon, M. ; Barreaux, S. (2018). « Lexique scientifique transdisciplinaire, terminologies et langues de spécialité en SHS », *in* Jacques, M.-P. ; Tutin, A. (dir.), *Lexique transversal et formules discursives des sciences humaines*, ISTE éditions, Londres, ISBN : 9781784054854 (papier), ISBN : 9781784064853 (ebook), p. 102-125.

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454

Kraif, O.; Diwersy S. (2012). « Le lexicoscope : un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques », dans *Actes de la conférence TALN 2012*, Grenoble, p. 399-406.

Kraif, O. (2016). « Le lexicoscope : un outil d'extraction des séquences phraséologiques basé sur des corpus arborés », *Cahiers de lexicologie : Revue internationale de lexicologie et lexicographie*, n° 108, p. 91-106, ISSN : 0007-9871.

Mathet, Y.; Widlöcher, A. (2016). « Évaluation des annotations : ses principes et ses pièges », dans *TAL*, 57(2), p. 73-98.

Paquot, M. (2010). "Academic vocabulary in learner writing : from extraction to analysis". *Research in corpus and discourse*, Bloomsbury Publishing, London, ISBN : 9781441114501.

Pecman, M. (2004). *Phraséologie contrastive anglais-français : analyse et traitement en vue de l'aide à la rédaction scientifique*, Thèse de doctorat, Université Nice Sophia Antipolis.

Silberztein, M. (2015). *La formalisation des langues : l'approche NooJ*, ISTE, Londres, ISBN : 978-1-78405-053-5 (papier), ISBN : 978-1-78406-053-4 (ebook).

Tutin, A. (2006). « Autour du lexique et de la phraséologie des écrits scientifiques », *Revue française de linguistique appliquée*, *12*(2), p. 5-14, ISSN : 1386-1204, e-ISSN : 1875-368X.

Tutin, A.; Jacques, M.-P. (2018). « Le lexique scientifique transdisciplinaire », *in* Jacques, M. P.; Tutin, A. (dir.), *Lexique transversal et formules discursives des sciences humaines*, ISTE Editions, Londres, p. 1-26, ISBN : 9781784054854 (papier), ISBN : 9781784064853 (ebook).

Published in : M. Mangeot et A. Tutin (dir.), *Lexique(s) et genre(s) textuel(s) : approches sur corpus*, Editions des archives contemporaines, ISBN : 9782813003454, doi : 10.17184/eac.9782813003454