SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100284

Relations syntaxiques entre lexiques terminologique et transdisciplinaire : analyse en texte intégral

Kister Laurence & Jacquey Evelyne

Atilf UMR 7118 CNRS/Université de Lorraine - Lexique - « Lexique et corpus »
Laurence.Kister@univ-lorraine.fr
Evelyne.Jacquey@atilf.fr

1 Introduction

L'extraction automatique de termes à partir de textes intégraux et l'évaluation des résultats de l'extraction sont régulièrement étudiées depuis les années 90 : (Daille 1996), (Bourigault *et al.* 2001 et 2004), (Drouin 2003 et 2004), (Bachimont *et al* 2005), (Kupść 2007), entre autres. La présence des termes dans les textes est une question sous-jacente à l'extraction de termes et au repérage de leurs variantes. En effet, la forme d'un terme peut faire l'objet de différents types de variations :

- des variations formelles, dues à la contextualisation, lorsque les termes ne se présentent pas dans le texte sous la même forme que dans les vocabulaires contrôlés du domaine que constituent les thesaurus, les nomenclatures, les lexiques et les terminologies. Ils peuvent faire l'objet d'une verbalisation et prendre une forme propre à un domaine ou un sous domaine (flexion, modifications adjectivales, ellipses, etc.),
- des variations de type homonymique, lorsqu'on est face à des lexicalisations susceptibles de produire des ambiguïtés, ce qui se produit le plus souvent pour des termes mono-lexicaux. On peut, par exemple, être en présence d'une ambiguïté entre un emploi terminologique et un emploi en langue générale, comme c'est le cas pour la forme *sujet*[1] qui peut faire l'objet dans un même document, d'un emploi terminologique propre au domaine des sciences du langage comme dans *le sujet de cette phrase* et un emploi en langue courante comme dans *le sujet que je développerai*,
- des variations sémantiques lorsque des ambiguïtés lexicales sémantiques résultent de l'existence de formes identiques pour désigner des concepts différents dans des domaines ou des sous-domaines proches. Le terme *discours*, par exemple, peut *désigner le langage mis en action* (synonyme de *parole*), *une unité égale ou supérieure à la phrase* (synonyme d'*énoncé*) en linguistique ou encore prendre un sens différent quand on adopte un point de vue rhétorique où il correspond à *une suite de développements oratoires destinés à persuader ou à émouvoir*[2].

La prise en compte des variantes permet de réduire le silence mais génère du bruit, ce qui alourdit la procédure de validation des candidats-termes. Le travail que nous présentons fait l'hypothèse que des lexèmes scientifiques transdisciplinaires scientifiques peuvent se trouver à proximité des termes du domaine et que, de ce fait, ils peuvent servir de marqueurs ou d'introducteurs de termes au sens où l'entendent (Frath *et al.* 2000) dont nous nous inspirons. Dans cette perspective, nous menons une expérience sur l'existence possible de relations syntaxiques entre les lexèmes scientifiques transdisciplinaires et les termes proches avec l'objectif de permettre une évaluation originale de la qualité des sorties des extracteurs. Nous nous focalisons sur un type particulier de lexèmes scientifiques transdisciplinaires introducteurs de termes : ceux qui fonctionnent comme des prédicats et qui prennent des termes pour dépendants syntaxiques. Les lexèmes scientifiques transdisciplinaires ont lorsqu'ils sont proches des termes une probabilité non négligeable d'être les recteurs de ces termes. Les lexèmes scientifiques transdisciplinaires recteurs constituent un argument en faveur de la validation du terme (53% des lexèmes scientifiques transdisciplinaires examinés sont recteurs d'un terme) : ils constituent un critère de différenciation des emplois terminologiques des termes et des emplois en langue courante de ces

Article available at http://www.shs-conferences.org or http://dx.doi.org/10.1051/shsconf/20120100284

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100284

mêmes termes. Cette première expérience porte sur des textes scientifiques et des textes de vulgarisation scientifique du domaine des sciences du langage.

2 Repérage des co-occurrences lexèmes transdisciplinaires-termes

2.1 Ressources textuelles, terminologiques et documentaires

Pour construire le matériel expérimental, nous utilisons trois types de ressources à partir desquelles nous confectionnons des échantillons destinés à tester la méthodologie avant de l'étendre à des corpus et des ressources terminologiques plus conséquents. Le matériel que nous avons constitué se compose de :

- un lexique scientifique transdisciplinaire composé d'une sélection de lexèmes scientifiques transdisciplinaires extraits des lexiques scientifiques transdisciplinaires mis au point par (Tutin 2007) et (Drouin 2007). Nous avons juxtaposé les lexiques que proposent ces deux auteurs puis nous les avons restreints aux entrées nominales (au nombre de 89 dont, par exemple, *cas, étude, travail, type*) et aux entrées verbales (au nombre de 60 dont, parmi d'autres, *agir, considérer, correspondre*). Pour cette phase de test, nous avons écarté les lexèmes de catégories adjectivales et adverbiales qui génèrent trop de bruit. A la suite de l'ébauche réalisée pour quelques lexèmes par (Drouin 2007), nous avons identifié parmi les lexèmes retenus certains lexèmes qui entrent dans la composition d'expressions plus ou moins figées. Nous considérons les expressions dans lesquelles apparaissent des lexèmes scientifiques transdisciplinaires comme des lexèmes scientifiques transdisciplinaires à part entière, car elles permettent d'introduire des termes du domaine de spécialité : *c'est le* ***cas*** *de, l'****étude*** *de, le* ***travail*** *sur/en, le* ***type*** *de, de type,* ***s'agir*** *de,* ***considérer*** *que,* ***correspondre*** *à*,
- une ressource terminologique qui provient de l'utilisation simultanée de deux sources : environ 6 100 entrées du vocabulaire de la linguistique de la base Francis[3] constitué par l'Inist et environ 1 200 termes de Thesaulangue,
- un corpus contrastif composé de 29 000 occurrences extraites des :
  - 300 000 occurrences du corpus scientifique Scientext[4] relatives aux sciences du langage, ce qui correspond à trois articles sceintifiques[5],
  - 170 000 occurrences d'un corpus de vulgarisation scientifique[6], mis à notre disposition par la revue Sciences Humaines sous licence CC, qui traite de sujets relatifs aux sciences du langage, ce qui correspond à trois articles de vulgarisation scientifique[7] touchant aux sciences du langage.

2.2 Repérage des co-occurrences

2.2.1 Extraction automatique de candidats-termes

L'expérience que nous avons réalisée a été effectuée en grande partie manuellement, cependant l'extraction des termes a fait l'objet d'une procédure automatique. Deux extracteurs mis à disposition de la communauté par leurs concepteurs ont été utilisés : *Acabit* (Daille 1996) et *TermoStat* (Drouin 2003). Ces deux extracteurs utilisent des étiqueteurs morphosyntaxiques : *Acabit* intègre un étiquetage avec *Brill*[8] et une lemmatisation grâce à *Flemm*, *TermoStat* utilise *TreeTagger*.

Les manières de procéder de ces extracteurs sont complémentaires. Les deux extraient des termes sur la base d'un coefficient statistique d'association entre les éléments des termes complexes et font usage des variantes syntagmatiques les plus connues (variations flexionnelles, introduction d'adjectifs, d'adverbes ou de groupes prépositionnels). Cependant, *TermoStat* apporte deux angles d'analyse supplémentaires : l'extraction des candidats-termes est mise en œuvre après la sélection des éléments les plus spécifiques de chaque texte analysé (calcul d'un coefficient de spécificité par confrontation avec vingt années du quotidien *Le Monde*) et l'extraction de candidats-termes mono-lexicaux en plus des candidats-termes

complexes. L'utilisation des deux extracteurs répond à deux problématiques fréquentes du TAL : la nécessité d'accroître le nombre de candidats-termes et l'amélioration de la qualité des candidats-termes par la prise en compte des candidats-termes complexes.

A partir des sorties des deux extracteurs, nous procédons à la mise en relief des candidats-termes repérés dans les textes intégraux sous forme d'un balisage XML :

<term key = 'acabit/termostat' id = ''> fautes d'orthographe </term>

Les textes ainsi balisés sont affichés dans une interface web afin de rendre l'évaluation des candidats-termes plus aisée et plus rapide. L'évaluation des candidats-termes est nécessaire au repérage de leurs co-occurrents, que ceux-ci appartiennent au lexique scientifique transdisciplinaire que nous utilisons pour cette expérience ou qu'ils relèvent de la langue de spécialité. L'enjeu du repérage des co-occurrences lexèmes transdisciplinaires-termes est l'identification des emplois terminologiques des termes. En effet, la proximité du langage de spécialité des sciences du langage et de la langue courante demande une attention particulière lors de l'évaluation des candidats-termes.

2.2.2 Evaluation manuelle par annotation des candidats-termes en texte intégral

La procédure d'annotation humaine a pour but de distinguer les candidats-termes qui font l'objet d'un emploi terminologique en sciences du langage, de ceux qui font l'objet d'un emploi en langue courante. La méthodologie, comme celle développée par (Delavigne 2001) sur des extraits de textes de vulgarisation scientifique dans le domaine de l'énergie nucléaire, procède par application de critères successifs aux candidats-termes. L'annotation par passages successifs permet d'éviter d'avoir à réaliser simultanément plusieurs annotations de natures différentes. Les couches successives d'annotation répondent chacune à une question élémentaire par deux valeurs : *valide* ou *non valide*. Les quatre étapes de l'annotation évaluent la validité syntagmatique du candidat-terme, son appartenance au lexique scientifique, son appartenance au lexique des sciences du langage et son emploi en tant que terme dans le contexte précis du document dans lequel il apparaît. Ainsi, à partir des candidats-termes détectés automatiquement, on procède progressivement à la sélection des termes relevant des sciences du langage qui font l'objet d'un emploi terminologique.

L'ensemble du corpus test a été traité par deux annotateurs. Le premier a systématiquement effectué une annotation intuitive tandis que le second a utilisé les ressources terminologiques disponibles. Il a vérifié la validité des candidats-termes par confrontation avec l'ensemble des vocabulaires scientifiques du portail Termsciences pour la couche-2 et avec les ressources spécifiques aux domaines des sciences du langage que sont le vocabulaire de la linguistique de Francis et Thesaulangue (inclus dans Termsciences) pour la couche_3 et la couche-4.

Le texte de départ ne contient que des pastilles vertes ( ) qui signalent les candidats-termes, ce qui signifie que tous les candidats-termes sont considérés comme *valides* par défaut. Lorsque l'annotateur choisit de rejeter un candidat-terme, il fait passer la pastille au rouge ( ) en cliquant dessus. Le passage du curseur sur les pastilles (↗) permet de mettre en évidence par coloration en bleu (**coloration en bleu**) le candidat-terme signalé par la pastille ce qui facilite la prise de décision, notamment dans le cas des candidats-termes poly-lexicaux.

Elle s'en distingue cependant, en [particulier], en privilégiant la ↗[**[combinaison de [[modèles]] théoriques] [variés]**], en favorisant les [[formalismes] [déclaratifs]], en permettant l' [élaboration graphique] des [[chaînes] de [traitement]] et en autorisant l' [[utilisation] de [tout] [↗ **[type]**] de **[**corpus**]**] XML. (Texte3)

Une procédure automatisée permet d'écarter les candidats-termes qui n'ont pas été validés, c'est-à-dire tous ceux qui ont été marqués d'une pastille rouge comme le montre le détail de l'annotation couche par couche proposé ci-dessous.

- **Couche_1 –** Exclusion des candidats-termes jugés non conformes d'un point de vue syntaxique.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100284

> Elle s'en distingue cependant, en [particulier], en privilégiant la [[combinaison de [[modèles]] théoriques] [variés]], en favorisant les [[formalismes] [déclaratifs]], en permettant l' [élaboration graphique] des [[chaînes] de [traitement]] et en autorisant l' [[utilisation] de [tout] [[type]] de [corpus]] XML.[9]

- La **couche_2 –** Exclusion des candidats-termes qui ne relèvent pas du lexique scientifique au sein des candidats-termes syntaxiquement conformes

> Elle s'en distingue cependant, en [particulier], en privilégiant la [combinaison de [[modèles] théoriques] variés], en favorisant les [[formalismes] déclaratifs], en permettant l' [élaboration graphique] des [[chaînes] de [traitement]] et en autorisant l' [utilisation] de [tout [[type]] de [corpus]] XML.

- La **couche_3** – Exclusion des candidats-termes qui ne relèvent pas de la langue de spécialité des sciences du langage au sein des candidats-termes qui appartiennent au lexique scientifique

> Elle s'en distingue cependant, en particulier, en privilégiant la [combinaison de [[modèles] théoriques] variés], en favorisant les [[formalismes] déclaratifs], en permettant l' [élaboration graphique] des [[chaînes] de [traitement]] et en autorisant l' [utilisation] de tout [[type] de [corpus]] XML.

- La **couche_4** – Exclusion des candidats-termes qui ne font pas l'objet d'un emploi terminologique au sein des candidats-termes qui font partie de la langue de spécialité des sciences du langage

> Elle s'en distingue cependant, en particulier, en privilégiant la [combinaison de [[modèles] théoriques] variés], en favorisant les [[formalismes] déclaratifs], en permettant l' [élaboration graphique] des [[chaînes] de [traitement]] et en autorisant l' [utilisation] de tout [[type] de [corpus]] XML.

- Résultat à l'issue de l'annotation

> Elle s'en distingue cependant, en particulier, en privilégiant la combinaison de modèles théoriques variés, en favorisant les formalismes déclaratifs, en permettant l'élaboration graphique des chaînes de traitement et en autorisant l'utilisation de tout type de [corpus] XML.

Cet exemple présente la particularité de conduire au même résultat à l'issue de la couche_2 et de la couche_3 : aucun candidat-terme n'a été rejeté au niveau de la couche_3. Pour cet exemple tous les candidats-termes qui appartiennent au vocabulaire scientifique sont des candidats-termes présents dans la langue de spécialité des sciences du langage. Il nous permet aussi de constater une forte diminution du nombre de candidats-termes après annotation puisque nous passons de 19 candidats-termes extraits automatiquement à 1 seul terme faisant l'objet d'un emploi terminologique validé par les annotateurs.

A l'image de cet exemple, nous constatons un écart important entre le nombre de candidats-termes syntaxiquement valides et celui des termes du domaine de spécialité des sciences du langage qui font effectivement l'objet d'un emploi terminologique dans les douments examinés.

	Scientext			Sciences Humaines		
	Texte1	Texte2	Texte3	Texte4	Texte5	Texte6
Candidats-termes extraits automatiquement	1601	1072	1121	1203	814	1146
Candidats-termes syntaxiquement valides	732	832	838	996	414	387
	45%	77%	74%	82%	50%	33%

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100284

Tableau 1 – Taux de candidats-termes syntaxiquement valides par rapport aux candidats-termes extraits automatiquement

	Scientext			Sciences Humaines		
	Texte1	Texte2	Texte3	Texte4	Texte5	Texte6
Candidats-termes syntaxiquement valides	732	832	838	996	814	387
Termes parmi les candidats-termes syntaxiquement valides	299	317	140	371	29	216
	40%	38%	16%	37%	7%	55%

Tableau 2 – Termes faisant l'objet d'un emploi terminologique par rapport aux candidats-termes syntaxiquement valides

A l'issue de l'annotation complète, nous obtenons un corpus de test constitué de six textes dans lesquels figurent les termes des sciences du langage qui font l'objet d'un emploi terminologique. Munis de ce corpus de travail, nous pouvons explorer notre hypothèse : évaluer le rôle d'introducteur de termes des lexiques scientifiques transdisciplinaires.

2.3 Répartition des co-occurrences lexèmes transdisciplinaires-termes

Cette étape du travail consiste à repérer les termes transdisciplinaires qui se trouvent à proximité d'un terme valide et qui entretiennent avec celui-ci une relation de dépendance syntaxique. Pour ce premier examen des relations entre termes et lexèmes scientifiques transdisciplinaires, nous définissons la *proximité* comme la co-occurrence intra-phrastique.

Comme nous l'avons signalé en 2.1 – Ressources textuelles, terminologiques et documentaires, les locutions ont été prises en compte. Ainsi, les locutions qui résultent de la co-occurrence *mettre* et de *jeu* ou d'*évidence* dans les locutions *mettre en jeu* et *mettre en évidence* sont considérées comme des lexèmes scientifiques transdisciplinaires alors que les lexiques scientifiques transdisciplinaires de (Drouin 2007) et (Tutin 2007) prennent en compte les deux lexèmes sceintifiques transdisciplinaires de manière automne :

> L'observateur trouvera donc matière dans des réalisations langagières de ce type à repérer et analyser des phénomènes que la figure [met en évidence]$_{\mathbf{LT}}$ de manière flagrante.
>
> La première catégorie des figures basée sur une référence plurielle [met en jeu]$_{\mathbf{LT}}$ l'appréhension distributive et collective de cette référence.
>
> Nous comprenons en effet sous le terme de synecdoque les relations tropiques qui [mettent en jeu]$_{\mathbf{LT}}$ le tout et la partie, le point d'unité entre les différentes catégories étant un rapport d'appartenance.

Cette étape a aussi permis de filtrer les constructions qui font intervenir une forme considérée par les lexiques scientifiques transdisciplinaires comme un *LT* alors que cette forme est un *T* dans les ressources terminologiques des sciences du langage. C'est, par exemple, le cas de *figure* qui apparaît dans les lexiques sceintifiques transdisciplinaires et, qui dans le contexte du Texte1 est un terme des sciences du langage qui correspond à *figure rhétorique*. Dans les exemples suivants, *figure* est un terme qui entretient une relation avec un lexème scientifique transdisciplinaire :

- *figure* entre dans une relation de type *LT de T* avec le lexème transdisiciplinaire *essentiel*

> En réalité, l'*essentiel*$_{\mathbf{LT}}$de la *figure*$_{\mathbf{T}}$ est ailleurs : dans la perception intuitive par l'*allocutaire*$_{\mathbf{T}}$ de la volonté d'exagération de la part du *locuteur*$_{\mathbf{T}}$

- *figure* se trouve dans une apposition et entretient une relation de type *LT de T* avec le lexème scientifique transdisciplinaire *effet* verbalisable en *l'effet des figures*

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100284

> L'effet$_{\mathbf{LT}}$, généralement reconnu aux figures$_{\mathbf{T}}$, est donc au minimum celui dû à une différence par rapport$_{\mathbf{LT}}$ à l'usage standard$_{\mathbf{T}}$.

Les relations syntaxiques entre termes et lexèmes scientifiques transdisciplinaires proches sont récurrentes dans le corpus expérimental (*cf.* Tableau 3 – Répartition des co-occurrences termes-lexèmes transdisciplinaires). Pour l'ensemble des textes, nous observons qu'environ un tiers des termes qui font l'objet d'un emploi terminologique entretiennent une relation de type syntaxique avec un des lexèmes scientifiques transdisciplinaires de la liste constituée pour cette expérience (36%).

	Scientext			Sciences Humaines			Total
	Texte1	Texte2	Texte3	Texte4	Texte5[10]	Texte6	
Nombre de T	299	317	140	371	29	216	1372
Co-occurrences T/LT	142	82	68	112	18	74	496
	47%	25%	48%	30%	62%	34%	36%

Tableau 3 – Répartition des co-occurrences termes-lexèmes transdisciplinaires

Parmi les co-occurrences lexèmes transdisciplinaires-termes, celles que nous examinons entretiennent une relation de dépendance syntaxique entre le lexème scientifique transdisciplinaire et le terme (Tableau 4 - Répartition des co-occurrences termes-lexèmes transdisciplinaires entretenant une relation de dépendance syntaxique).

	Scientext			Sciences Humaines			
	Texte1	Texte2	Texte3	Texte4	Texte5	Texte6	Total
Co-occurrences T/LT	142	82	68	112	18	74	496
Dépendances syntaxiques T/LT	89	38	46	48	5	38	264
	62%	46%	80%	38%	27%	51%	53%
Dépendances par corpus	59%			44%			

Tableau 4 – Répartition des co-occurrences termes-lexèmes transdisciplinaires entretenant une relation de dépendance syntaxique

Les premiers résultats montrent que les trois textes scientifiques examinés comportent une proportion légèrement plus importante de co-occurrents lexèmes transdisciplinaires-termes reposant sur une relation syntaxique que les trois textes de vulgarisation (59% contre 44%). Les proportions restent cependant trop proches pour tirer une conclusion quant à la répartition et l'utilisation des termes et des lexèmes scientifiques transdisciplinaires en fonction du type de document : les chiffres plaident plutôt en faveur de l'absence de discrimination à partir de ce critère et l'existence d'un rôle d'introducteur de terme pour les lexèmes scientifiques transdisciplinaires, tant en texte scientifique qu'en texte de vulgarisation scientifique. Ces résultats montrent aussi qu'il ne s'agit pas de tendances spécifiques aux types de textes puisque les résultats sont variables en fonction des auteurs dans les deux types de textes : de 25% à 48% pour les textes scientifiques et de 30% à 62% pour les textes de vulgarisation.

3 Analyse des co-occurrences *lexèmes transdisciplinaires-termes*

3.1 Répartition des co-occurrences

Les premiers résultats de cette phase expérimentale montrent un comportement intéressant sur le plan de l'articulation entre les termes et la liste de lexèmes scientifiques transdisciplinaires constituée à partir des lexiques scientifiques transdisciplinaires de (Drouin 2007) et (Tutin 2007). Le nombre de co-occurrences lexèmes transdisciplinaires-termes est de 115 pour les lexèmes scientifiques transdisciplinaires nominaux et de 113 pour les lexèmes scientifiques transdisciplinaires verbaux. Si on s'intéresse au nombre de lexèmes scientifiques transdisciplinaires co-occurrents de termes pour lesquels on est en présence d'une dépendance syntaxique, on observe 44 lexèmes scientifiques transdisciplinaires nominaux (49% des

lexèmes scientifiques transdisciplinaires nominaux) et 28 lexèmes scientifiques transdisciplinaires verbaux (46% des lexèmes scientifiques transdisciplinaires verbaux).

Certains lexèmes scientifiques transdisciplinaires, tant nominaux que verbaux, apparaissent dans un seul type de texte tandis que d'autres sont présents dans les deux types de texte[11].

	Scientext	Sciences Humaines	Scientext et Sciences Humaines	
Lexèmes transdisciplinaires nominaux	21	10	13	44
	48%	22%	30%	100%
Lexèmes transdisciplinaires verbaux	10	5	13	28
	36%	17%	47%	100%

Tableau 5 – Répartition des lexèmes scientifiques transdisciplinaires qui entretiennent une relation syntaxique avec des termes en fonction du type de texte

L'utilisation des lexèmes scientifiques transdisciplinaires n'apparaît pas comme une caractéristique propre de l'un ou l'autre des types de textes. Pour ce qui de la répartition des différents lexèmes par type de texte, nous pouvons constater qu'ils sont relativement nombreux à apparaître dans les deux types de documents : 21 lexèmes scientifiques transdisciplinaires sont présents dans les deux types de textes, 31 n'apparaissent qu'au niveau des textes scientifiques et 15 au niveau des textes de vulgarisation. La répartition des lexèmes scientifiques transdisciplinaires ne permet pas de tirer de conclusion plus précise en raison de la différence de taille des deux corpus (300 000 occurrences contre 170 000).

3.2 Constructions syntaxiques associant termes et lexèmes transdiciplinaires

Les 44 lexèmes scientifiques transdisciplinaires nominaux et les 28 lexèmes scientifiques transdisciplinaires verbaux produisent respectivement 76 et 102 co-occurrences pour lesquelles le lexème scientifiques transdisciplinaire entretient une relation syntaxique avec un terme. Afin de catégoriser les relations, nous opérons une première répartition des co-occurrences en fonction du type de dépendance qui unit le terme et le lexème scientifique transdisciplinaire : la relation directe entre le terme et son recteur correspond à ce que nous désignons par *dépendance directe* et la dépendance entre un syntagme ou une proposition contenant un terme et le recteur de la structure syntaxique complexe que nous considérons comme une *dépendance indirecte*.

Afin de pouvoir entreprendre une analyse des relations entre les lexèmes scientifiques transdisciplinaires et les termes et vérifier si l'hypothèse selon laquelle ces lexèmes scientifiques transdisciplinaires peuvent constituer des marqueurs/des indices de la présence de termes est valide, nous effectuons des relevés systématiques et alimentons des tableaux d'analyse destinés à repérer les relations les plus fréquentes afin de prévoir une automatisation du repérage des termes à partir des lexèmes scientifiques transdisciplinaires.

3.2.1 Relations directes

Nous parlons, par exemple, de relation directe lorsque le terme entretient une relation verbale directe avec le lexème scientifique transdisciplinaire, comme c'est le cas pour *T = dyslexie* et *LT = être* dans ***être** un trouble de* dont *T* est argument. Il s'agit pour l'exemple ci-dessous d'une relation sujet :

La dyslexie$_{\mathbf{T}}$ est$_{\mathbf{LT}}$ un trouble spécifique de l'acquisition de la lecture

Un cas de figure similaire se produit quand le terme entretient une relation nominale telle celle qui s'établit entre *T = lecture* et *LT = acquisition* :

La dyslexie est un trouble spécifique de l'acquisition$_{\mathbf{LT}}$ de la lecture$_{\mathbf{T}}$

Les **lexèmes scientifiques transdisciplinaires nominaux** entretiennent avec les termes des relations syntaxiques pour lesquelles *T* est complément d'un *LT nominal recteur* (concept$_{LT}$ de dyslexie$_{T}$ -

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100284

analyse$_{LT}$ des sons$_{T}$ du langage$_{T}$), *T* est sujet et *LT* est attribut (la lecture$_{T}$ n'est pas simplement une activité$_{LT}$ visuelle) ou encore *T* est objet et *LT* est sujet (que ces derniers$_{LT}$ soient ou non spécifiques au langage$_{T}$).

Les **lexèmes scientifiques transdisciplinaire verbaux** et les termes entrent, quant à eux, dans des relations syntaxiques où *T* est sujet de *LT* (la dyslexie$_{T}$ concerne$_{LT}$ 5% des enfants) et où *T* est un objet qui dépend d'un *LT* verbal ((l'intelligence) permet de définir$_{LT}$ la dyslexie$_{T}$ - on utilise$_{LT}$ pour l'apprentissage de la lecture une écriture$_{T}$).

	Position de T	Exemple	Recteur de T
Lexèmes transdisciplinaires nominaux	Comp N	concept$_{\mathbf{LT}}$ de dyslexie$_{\mathbf{T}}$	**concept**
	Comp N (T de T)	analyse$_{\mathbf{LT}}$ des sons$_{\mathbf{T}}$ du langage$_{\mathbf{T}}$	**analyse**
	Comp N (N de T)	étude$_{\mathbf{LT}}$ sur les troubles$_{\mathbf{N}}$ du langage$_{\mathbf{T}}$	**étude**
	Sujet	la lecture$_{\mathbf{T}}$ n'est pas simplement une activité$_{\mathbf{LT}}$ visuelle	être (attribut)
	Comp à	que ces derniers$_{\mathbf{LT}}$ soient ou non spécifiques au langage$_{T}$	être spécifique
Lexèmes transdisciplinaires verbaux	Comp	(l'intelligence) permet de définir$_{\mathbf{LT}}$ la [dyslexie]$_{\mathbf{T}}$	**définir**
	Sujet	la dyslexie$_{\mathbf{T}}$ concerne$_{\mathbf{LT}}$ 5 % des enfants	**concerner**
	Comp (pour)	on utilise$_{\mathbf{LT}}$ pour l'apprentissage de la lecture$_{\mathbf{T}}$ une écriture$_{\mathbf{T}}$	**utiliser**

Tableau 6 – Exemples de relations syntaxiques directes entre lexèmes scientifiques transdisciplinaires et termes

3.2.2 Relations indirectes

La relation indirecte concerne, par exemple, un terme qui entretient une relation verbale indirecte de type attribut comme *T = lecture* avec *T = dyslexie* par l'intermédiaire du lexème scientifique transdisciplinaire *LT = **être*** de ***être** un trouble*.

La dyslexie$_{\mathbf{T}}$ est$_{\mathbf{LT}}$ un trouble spécifique de l'acquisition de la lecture$_{\mathbf{T}}$

Les **lexèmes scientifiques transdisciplinaires nominaux** peuvent être recteurs du terme lorsque *T* est un complément prépositionnel lui-même inclus dans un complément prépositionnel (ces deux types$_{LT}$ de troubles de la lecture$_{T}$), lorsque *T* est inclus dans un comparatif (admettre qu'une activité$_{LT}$ aussi subtile que la lecture$_{T}$), ou encore lorsque *T* appartient à une relative (les résultats d'études$_{LT}$ dans lesquelles on a suivi les mêmes enfants avant et après l'[apprentissage de la lecture]$_{T}$ indiquent que...).

Ils peuvent aussi participer à des structures pour lesquelles *LT* entretient une relation avec *T* mais n'est pas son recteur. C'est le cas lorsque *T* est régi par *verbe + avoir* (ces difficultés$_{LT}$ peuvent avoir des origines diverses comme une mauvaise maîtrise de la langue$_{T}$), lorsque *T* est régi par un auxiliaire et inclus dans un comparatif (un autre point $_{LT}$ crucial est que les [études longitudinales]$_{T}$) et lorsque *T* est adjoint et est régi par un verbe ou un auxiliaire (ce déficit se manifeste plus ou moins fortement en fonction$_{LT}$ de la transparence de l'orthographe$_{T}$ - dans ce contexte$_{LT}$, une explication plausible est que pour mettre en relation les graphèmes$_{T}$ avec les phonèmes correspondants…).

Les **lexèmes scientifiques transdisciplinaires verbaux** entretiennent des relations syntaxiques où *LT* est recteur de *T* comme c'est le cas quand *T* est sujet de *LT* (que le terme$_{T}$ dyslexie$_{T}$ ne devienne$_{LT}$ le fourre-tout de l'échec scolaire - l'identification des mots$_{\mathbf{T}}$ peut être obtenue$_{\mathbf{LT}}$ soit par une procédure globale) et quand *T* est complément de *LT* (quand ils doivent$_{LT}$ lire des mots$_{T}$ nouveaux - qui permet$_{LT}$ de définir la dyslexie$_{T}$). A côté des relations pour lesquelles *LT* est recteur de *T*, il existe des relations pour lesquelles *LT* n'est pas recteur de *T*. *T* peut alors être complément d'un N lui même complément de *LT* le tout est alors régi par un verbe (…qui l'a utilisé en 1892 pour décrire$_{\mathbf{LT}}$ les troubles de la lecture$_{\mathbf{T}}$ - quelques rares enquêtes épidémiologiques permettent$_{\mathbf{LT}}$ de penser que la dyslexie$_{T}$ concerne 5 % des enfants).

Pour chaque terme qui entretient une relation de proximité avec un lexème scientifique transdisciplinaire, nous cherchons la présence d'une relation de dépendance directe puis d'une relation indirecte. Les relevés

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100284

effectués sur le corpus de textes qui nous sert d'échantillon font apparaître des régularités encourageantes comme le montrent les exemples de relations directes et indirectes proposées ci-dessus.

	Position de T	Exemple	Recteur de T
Lexèmes transdisciplinaires nominaux	Comp N	ces deux types$_{LT}$ de troubles de la lecture$_{T}$	**Type**
	Comp (exemple)	ces difficultés$_{LT}$ peuvent avoir des origines diverses comme une mauvaise maîtrise de la langue$_{T}$	avoir V support des origines
	Comp (apposition)	la compréhension$_{LT}$ d'un texte, finalité de la lecture$_{T}$	**compréhension**
	Comp (comparatif)	admettre qu'une activité$_{LT}$ aussi subtile que la lecture$_{T}$	**activité**
	Comp (que)	Un autre point$_{LT}$ crucial est que les [études longitudinales]$_{T}$	être
	Adjoint (en fonction de)	Ce déficit se manifeste plus ou moins fortement en fonction$_{LT}$ de la transparence de l'orthographe$_{T}$	se manifester
	Comp (relative)	les résultats d'études$_{LT}$ dans lesquelles on a suivi les mêmes enfants avant et après l' [apprentissage de la lecture]$_{T}$ indiquent que	**étude**
	Adjoint (pour)	Dans ce contexte$_{LT}$, une explication plausible est que pour mettre en relation les graphèmes$_{T}$ avec les phonèmes correspondants	**être**
	Comp	Il permet également de comprendre le retard de l'écriture$_{T}$ sur la lecture, conséquence$_{LT}$ de l'asymétrie des relations	**comprendre**
Lexèmes transdisciplinaires verbaux	Sujet	que le terme$_{T}$ dyslexie$_{T}$ ne devienne$_{LT}$ le fourre-tout de l'échec scolaire	**devenir**
	Comp	qui l'a utilisé en 1892 pour décrire$_{LT}$ les troubles de la lecture$_{T}$	utiliser
		quand ils doivent$_{LT}$ lire des mots$_{T}$ nouveaux	**devoir**
	Comp (infinitif)	qui permet$_{LT}$ de définir la dyslexie$_{T}$	**permettre**
	Comp (que)	Quelques rares enquêtes épidémiologiques permettent$_{LT}$ de penser que la dyslexie$_{T}$ concerne 5 % des enfants	penser
	Sujet	l'identification des mots$_{T}$ peut être obtenue$_{LT}$ soit par une procédure globale	être obtenu
	Relative	l'enfant qui fait$_{LT}$ quelques fautes$_{N}$ d'orthographe$_{T}$	enfant

Tableau 7 – Exemples de relations syntaxiques indirectes entre lexèmes transdisciplinaires et termes

4 En guise de conclusion…

Cette première expérience, bien que restreinte à six documents et à la juxtaposition d'extraits de lexiques scientifiques transdisciplinaires, montre un rôle certain des lexèmes scientifiques transdisciplinaires en tant qu'introducteur de termes. En effet, lorsque nous sommes en présence d'un lexème sceintifique transdisciplinaire et d'un terme dans une même phrase, dans près de 50% des cas, le lexème scientifique transdisciplinaire entretient une relation syntaxique avec le terme et joue un rôle d'indice de l'emploi terminologique de celui-ci. Afin de parfaire les résultats et d'estimer plus précisément le rôle des lexèmes scientifiques transdisciplinaires dans l'introduction de termes du domaine, il est nécessaire d'étendre l'analyse à un corpus plus vaste : une campagne d'annotation et d'analyse d'une soixantaine de textes scientifiques en sciences du langage extraits de Scientext est en cours[12]. Le choix de nous focaliser sur les textes scientifiques se justifie par la répartition relativement similaire des résultats que nous avons observée pour les textes scientifiques et les textes de vulgarisation scientifique. La mise à disposition par Scientext d'un nombre relativement conséquent de documents balisés est un second argument en faveur de notre décision et participe à la mutualisation des résultats de la recherche en sciences humaines.

SHS Web of Conferences 1 (2012)
DOI 10.1051/shsconf/20120100284

La campagne d'annotation de texte scientifique sera suivie, à l'image de ce que nous avons fait pour cette première expérience, d'un relevé systématique des phrases qui comportent des lexèmes scientifiques transdisciplinaires et/ou des termes faisant l'objet d'un emploi terminologique dans le contexte des textes examinés. Ces relevés permettront d'évaluer la nature et le nombre des lexèmes scientifiques transdisciplinaires, des termes de la langue de spécialité des sciences du langage et des relations syntaxiques qui s'établissent entre lexèmes scientifiques transdisciplinaires et termes. L'analyse en termes de relations syntaxiques présente l'avantage de pouvoir être informatisée relativement aisément ce qui peut alléger le travail manuel et permettre un repérage assisté des emplois terminologiques des termes en texte intégral.

Pour apporter une réponse complète à la question que nous avons posée sous forme d'hypothèse, il nous faudra aussi compléter les comptages afin de proposer des estimations fiables. Nous devons à moyen terme examiner la manière dont les termes sont introduits quand ils ne sont pas à proximité d'un lexème scientifique transdisciplinaire et le rôle que peuvent jouer les lexèmes scientifiques transdisciplinaires qui ne sont pas introducteurs de termes.

Références bibliographiques

Bachimont B, Baneyx A, Malaisé V, Charlet J et Zweigenbaum P. (2005). Synergie entre analyse distributionnelle et patrons lexico-syntaxiques pour la construction d'ontologies différentielles, *TIA*, Rouen.

Bourigault D, Aussenac-Gilles N et Charlet J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. M. Slodzian (ed*), Revue d'Intelligence Artificielle, Numéro spécial sur les techniques informatiques de structuration de terminologies*, Hermès, 18(1), 87-110.

Bourigault D, Jacquemin C et L'Homme MC. (2001). *Recent Advances in Computational Terminology*. Amsterdam-Philadelphie : John Benjamins

Daille B. (1996). ACABIT : une maquette d'aide à la construction automatique de banques terminologiques, in Clas A, Thoiron P et Béjoint H, (eds.), *Lexicomatique et Dictionnairitique*, FMA, Beyrouth, 123-136.

Delavigne V. (2001). Repérage de termes dans un corpus de vulgarisation : aspects méthodologiques, *TIA*, Nancy, 33-43.

Drouin P. (2007). Identification automatique du lexique scientifique transdisciplinaire, *Revue Française de linguistique appliquée*, 12(2), 45-64.

Drouin P. (2004). Acquisition des termes simples fondée sur les pivots lexicaux spécialisés. *TIA,* Strasbourg.

Drouin P. (2003). Term extraction using on-technical corpora as a point of leverage. *Terminology*, 9(1), 99-117

Dubois J, M. Giacomo, L. Guespin et C. Marcellesi (2001). *Dictionnaire de linguistique et des sciences du langage*, Paris : Larousse, 1ère édition 1994.

Frath P Oueslati R et Rousselot F. (2000). Identification de relations sémantique par repérage et analyse de cooccurrences de signes linguistiques, in Charlet J, Zacklad M, Kassel G et Bourrigault (eds), *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, Eyrolles.

Journet N. (1999). Profession linguiste, in *Le Langage*, Sciences Humaines, Hors-série n° 27.

Kupść, A. (2007). Extraction automatique de termes à partir de textes polonais, *TALN*, Toulouse.

Lecolle, M. (2000). Figures et références plurielle en corpus journalistique, in *Cahiers de Grammaire*, n° 25.

Moser P. K., Trout J.D. et Mulder D. (1999). La raison, l'expérience et la confiance, in *La dynamique des savoirs*, Sciences Humaines, Hors-série n° 24.

Payre-Ficout, C. et Chevrot J.P. (2001). La forme contre l'usage: étude exploratoire de l'acquisition du prétérit anglais, in Acquisition et enseignement de la morphographie, *LIDIL*, n°30.

Wildöcher A. et Bilhaut F. (2007). La plate-forme LinguaStream, in *Autour des langues et du langage: perspective pluridisciplinaire,* Presses Universitaires de Grenoble, Grenoble, 447-454.

Sprenger-Charolles L. (2003). La dyslexie repensée, in *Sciences humaine*, Mensuel n° 134.

Tutin A. (2007). Lexique et écrits scientifiques. *Revue française de linguistique appliquée*, XII(2).

[1] Le terme *sujet* apparaît dans le micro-thesaurus de la *syntaxe* de Thesaulangue - le thesaurus des sciences du langage mis au point et maintenu par le centre de documentation de l'Atilf. Thesaulangue est accessible pour la communauté par le portail TermSciences de l'Inist.

[2] Les éléments de définition sont extraits de Dubois J, M. Giacomo, L. Guespin et C. Marcellesi. (2001). *Dictionnaire de linguistique et des sciences du langage*, Paris : Larousse, 1ère édition 1994.

[3] La base FRANCIS signale près de 2,5 millions de références en sciences humaines et sociales. 60 000 nouvelles références issues de l'analyse de plus de 2000 revues scientifiques internationales sont ajoutées par an (*chiffres de 2009*).

[4] Le projet ANR Scientext met à la disposition de la communauté scientifique un corpus d'écrits scientifiques consultable en ligne d'environ 4,4 millions de mots en français et 33 millions de mots en anglais. Le corpus a été analysé syntaxiquement et annoté au plan structurel en suivant les recommandations de la Text Encoding Initiative (TEI Lite), en isolant les différentes parties textuelles de l'article : résumé, introduction, corps du texte, conclusion, remerciements, notes de bas de page, bibliographie, annexes, titres.

[5] Références des textes scientifiques provenant du corpus Scientext :

- (Texte1) M. Lecolle, 2000, Figures et références plurielle en corpus journalistique, *Cahiers de Grammaire*, n° 25.
- (Texte2) C. Payre-Ficout et J.P. Chevrot, 2001, La forme contre l'usage: étude exploratoire de l'acquisition du prétérit anglais, Acquisition et enseignement de la morphographie, *LIDIL*, n°30.
- (Texte3) A. Wildöcher et F. Bilhaut, 2007, La plate-forme LinguaStream, *Autour des langues et du langage: perspective pluridisciplinaire*, Presses Universitaires de Grenoble, Grenoble, 447-454.

[6] La mise au format TEI du corpus de vulgarisation fourni par la revue Sciences Humaines a été réalisée à l'Atilf, par l'équipe *Ressources et normalisation*.

[7] Références des textes de vulgarisation scientifique provenant de la revue Sciences Humaines :

- (Texte4) N. Journet, 1999, Profession linguiste, *Sciences Humaines*, Le Langage - Hors-série n° 27.
- (Texte5) P. K. Moser, J.D. Trout et D. Mulder, 1999, La raison, l'expérience et la confiance, *Sciences Humaines*, *La dynamique des savoirs* - Hors-série n° 24.
- (Texte6) L. Sprenger-Charolles, 2003, La dyslexie repensée, *Sciences humaine*, Mensuel n° 134.

[8] Une procédure intégrée de traitement d'un texte par *Acabit* en utilisant les fichiers d'entraînement de *Brill* réalisés à partir de Frantext a été mise au point par l'équipe à *Soutien technique à la recherche* de l'Atilf.

[9] Cet exemple est extrait du Texte3.

[10] Le (Texte5) se distingue par un score extrêmement faible car seul un paragraphe concerne réellement la linguistique alors qu'il a été identifié comme un texte de linguistique par le moteur de recherche du site de la revue Sciences Humaines à partir des requêtes que nous avons effectuées.

[11] Le tableau ci-dessous répertorie les lexèmes scientifiques transdisciplinaires en fonction de la catégorie de texte dans laquelle ils apparaissent.

	Scientext	Sciences Humaines	Scientext et Sciences Humaines
Lexèmes transdisciplinaires nominaux	analyse auteur caractère cas choix compréhension contexte différence difficulté document effet ensemble essentiel forme influence jeu lieu modèle objet règle	activité article caractéristique dernier état lien maintien succès totalité unité	approche élément étude exemple fait fonction nombre part point rapport système travail type
Lexèmes transdisciplinaires verbaux	agir décrire écarter manquer négliger ouvrir prendre représenter subdiviser suggérer	concerner dépasser lier proposer sembler	considérer constituer correspondre définir donner faire mettre montrer permettre pouvoir présenter trouver utiliser

[12] Cette campagne d'annotation est réalisée dans le cadre du projet ASTTIC (Annotation Sémantique et Terminologique de Textes pour leur Indexation et leur Catégorisation) – Axe 2 : Langues, textes, documents - pour lequel nous bénéficions d'un financement de la MSH-Lorraine.