Extraction de collocations à partir du champ syntagme du TLFi : application aux noms transdisciplinaires des écrits scientifiques

Veronika Lux-Pogodalla, Agnès Tutin

Extraction de collocations à partir du champ syntagme du *TLFi* : application aux noms transdisciplinaires des écrits scientifiques

Veronika Lux-Pogodalla (1) veronika.lux@inist.fr

Agnès Tutin (2) agnes.tutin@u-grenoble3.fr

(1) Inist-CNRS, 2, allée de Brabois, 54514 Vandoeuvre-lès-Nancy

(2) LIDILEM, Université Grenoble 3, UFR des sciences du langage, BP25, 380440 Grenoble cedex 9

Mots-clés : lexique scientifique général, collocations, extraction de collocations, aide à la rédaction

Keywords: general scientific lexicon, collocations, collocation extraction, writing help

Résumé : Le *TLFi* est une source de données lexicales extrêmement riche et relativement peu exploitée. Dans cet article, nous souhaitons évaluer dans quelle mesure il est possible d'extraire semi-automatiquement un sous-ensemble de collocations (champ « syntagme » du *TLFi*) liées au lexique général des écrits scientifiques, ce lexique transdisciplinaire qui renvoie à la description, au processus et au raisonnement des activités scientifiques. Nous voudrions déterminer comment et dans quelle mesure on peut tirer parti de cette ressource électronique en la combinant avec des informations automatiquement extraites de corpus à l'aide d'un analyseur syntaxique.

Abstract : The *TLFi* is a rich and underexploited source of lexical data. Here, we want to evaluate if a list of collocations can be semi-automatically extracted (using the "syntagme" field of the TFLi), that are related to the general lexicon of scientific writing. Items in this lexicon are used for the descriptions, the processes and the reasonings of scientific activities. We want to determine how and to which extend this electronic resource can be used for our aim, combining data extracted from the *TLFi* with data automatically extracted from text corpora.

1 Problématique

Le *TLFi* est une source de données lexicales extrêmement riche qui, pour certain de ses champs, a de façon surprenante été assez peu exploitée. Le champ syntagme, qui nous intéresse particulièrement, a, à notre connaissance, peu fait l'objet d'études systématiques, hormis celles de [Hausmann, 1995]. Dans cet article, nous souhaitons évaluer dans quelle mesure il est possible d'extraire semi-automatiquement un sous-ensemble de collocations liées au lexique général des écrits scientifiques, ce lexique transdisciplinaire qui renvoie à la description, au processus et au raisonnement des activités scientifiques. Nous voudrions déterminer comment et dans quelle mesure on peut tirer parti de cette ressource électronique en la combinant avec des informations extraites de corpus à l'aide d'un analyseur syntaxique. Nous examinerons ainsi les collocations de type V-N (*faire l'hypothèse, vérifier l'hypothèse, mener une étude*) et Adj-N (*une hypothèse pertinente, des résultats encourageants*) extraites semi-automatiquement du champ syntagme du *TLFi* et les extractions effectuées entièrement automatiquement à partir d'un corpus d'écrits scientifiques de 2 millions de mots analysé à l'aide du logiciel Syntex [Bourigault, 2007]. La façon dont on peut combiner ces deux ressources sera évaluée par des linguistes s'intéressant à la problématique du lexique des écrits scientifiques.

2 Les collocations de langue scientifique générale

Dans le cadre de cette expérimentation, nous nous intéressons aux collocations transdisciplinaires des écrits scientifiques. Les collocations étant une notion à géométrie variable, il convient bien entendu d'en préciser les contours. Notre approche des collocations se situe dans la lignée de Mel'čuk et Hausmann (par exemple, [Hausmann, 1989], [Mel'čuk, 1998]). Ce sont pour nous des expressions linguistiques composées de deux éléments linguistiques, apparaissant fréquemment en cooccurrence et entretenant une relation syntaxique, et dont l'un des éléments, la base conserve son sens habituel, alors que le collocatif apparaît moins prédictible. Dans *résultats encourageants,* le mot *résultats* aurait ainsi la fonction de base, alors que *encourageants* ferait office de collocatif, étant conditionné par le mot *résultats*. Nous nous intéressons ici particulièrement aux collocations apparaissant dans les écrits scientifiques et qui sont emblématiques de ce genre, renvoyant à la description de l'activité scientifique, aux résultats, aux évaluations, au raisonnement mis en jeu dans les écrits de ce type. Ce lexique, qui transcende en grande partie les disciplines (il ne renvoie pas à la terminologie du domaine), comporte des expressions comme *faire une hypothèse, rejeter une hypothèse, approche traditionnelle, étude théorique, thèse classique* … . Notre objectif à plus long terme est de proposer un inventaire exhaustif et une description sémantique de ce lexique pour des applications d'aide à la rédaction en langue étrangère. Notre expérimentation vise à évaluer l'emploi du *TLFi* comme source possible pour l'acquisition de ce lexique.

3. Extraction des collocations du lexique transdisciplinaire du *TLFi*

Le *TLF* est souvent considéré comme un dictionnaire de langue littéraire. La langue scientifique et technique n'y a cependant pas été négligée et nous pensons que le *TLFi* peut en partie servir de base pour la constitution d'un dictionnaire des collocations de la langue scientifique générale, même si l'état de langue décrit y est un peu ancien et le lexique extrait doit être en partie actualisé.
Un avantage considérable du *TLF* est la possibilité d'extraire des éléments de champs ciblés à partir de la version informatisée balisée, procédure déjà adoptée pour des dictionnaires bilingues (Cf. par exemple [Fontenelle, 1997]). En outre, par rapport à d'autres dictionnaires de langue comme le *Petit Robert*, le *TLF* apparaît très bien structuré dans le traitement des collocations. Tout d'abord, le concept de « syntagme » (à peu près équivalent à notre notion de collocation) est différencié de la notion de locution et fait en principe l'objet d'un traitement spécifique[1], même si dans les faits, quelques incohérences demeurent [Henry, 1995]. L'informatisation du dictionnaire permet d'avoir accès à la « richesse de la face cachée » du *TLF*, comme le décrit [Haussmann, 1995] (p. 38), en accédant aux collocations à partir de certains champs de l'article : (a) les éléments retenus dans la rubrique « SYNT. » (pour « syntagme »), une rubrique spécifiquement dédiée aux informations de cooccurrences (Cf. (1) ci-dessous),

> (1) *TLFI* s.v. *conclusion* II.A[2] → le champ « SYNT »
> SYNT. (très fréq.). *Conclusion audacieuse, bonne, erronée, hardie, juste, nulle, prématurée; les conclusions qui se dégagent de ces travaux; exposer la conclusion de ses réflexions; aboutir, arriver, être conduit à certaines conclusions, aux conclusions suivantes; affirmer, confirmer, infirmer une conclusion; établir, formuler une conclusion inattaquable; amener qqn à ses conclusions.*

(b) les « collocations enchaînées » qui suivent la définition, mais ne sont pas glosées ou définies, (c) les « collocations définies », qui ne sont introduites par aucun indicateur, (d) les fausses locutions, qui sont des sous-vedettes[3], introduites par l'indicateur « Loc. », qui sont en réalité des collocations. Haussmann signale également d'autres champs où les collocations apparaissent également de façon plus diffuse, et où l'extraction automatique n'apparaît pas envisageable en l'état : (e) les collocations dans les citations, détachées ou enchaînées ; (f) les collocations dans les définitions, où la collocation doit être reconstruite ; (g) les collocations synonymiques ou antonymiques.
Dans le *TLFi*, on peut extraire les collocations du champ « syntagmes»[4] qui correspond à peu près aux champs (a)-(d) du *TLF* « papier ». Dans cette version électronique, on pourra également effectuer une recherche sur tous les articles à partir de la recherche assistée ou de la recherche complexe, en extrayant tous les articles qui contiennent un mot donné dans le champ syntagme. Il est même possible de préciser dans la recherche complexe dans quel type de champ on veut extraire le syntagme : paragraphe syntagme (rubrique dédiée), syntagme défini, ou syntagme enchaîné (que l'indicateur « Loc. » apparaisse ou non). Sur l'interface en ligne, il est ensuite possible d'afficher tous les champs contenant les syntagmes, sans passer par l'affichage de l'article complet.

Pour cette communication, nous avons obtenu la liste des syntagmes comprenant 90 noms considérés comme relevant du lexique général (*hypothèse, cas, données, thèse, approche* …)[5] sous forme de texte balisé XML[6] dont nous avons extrait un sous-ensemble de syntagmes sous les vedettes verbales et adjectivales. Par exemple, sous la vedette verbale MENER, on repère des syntagmes comportant les noms transdisciplinaires *analyse, étude* et *recherches*, comme on peut le voir dans la Figure 1.

```
<article>
  <vedette>MENER, verbe trans.</vedette>
    <occurrences>
```

[1] [Henry, 1995] mentionne (p.107) l'existence de définitions précises pour les notions de « syntagme », « locution » et « phrase figée ayant valeur de vérité générale » dans le *Cahier des normes* utilisé en interne pour la rédaction, tout en signalant quelques incohérences dans le traitement des articles.
[2] La définition principale donnée pour cette acception est la suivante : « Proposition tirée des données de l'observation ou d'un raisonnement. »
[3] Dans le *TLFi*, le terme sous-vedette a une extension plus étroite. Il semble surtout renvoyer à des mots composés très figés, souvent réunis par des traits d'union.
[4] Dans le *TLF*, on peut postuler les degrés de figement suivants pour les unités polylexicales (du plus figé au moins figé) : vedettes (ex : *pied à terre*), sous-vedette (ex : *maison-témoin*), syntagme défini avec marqueur « Loc. » (ex : *casser la tête de qqun*), syntagme défini (ex : *maison seigneuriale*), les syntagmes enchaînés (ex : *crainte, peur irraisonnée*).
[5] Voir liste en annexe.
[6] Un grand merci à Etienne Petitjean de nous avoir fourni cette liste.

```
        <occurrence>
          <mot>analyse</mot>
            <syntagme>Mener une action, une analyse, une étude, une politique, des
            recherches.</syntagme>
        </occurrence>
        <occurrence>
          <mot>étude</mot>
            <syntagme>Mener une action, une analyse, une étude, une politique, des
            recherches.</syntagme>
        </occurrence>
        <occurrence>
          <mot>recherches</mot>
            <syntagme>Mener une action, une analyse, une étude, une politique, des
            recherches.</syntagme>
        </occurrence>
    </occurrences>
</article>
```

Figure 1 : Extrait du *TLF* de syntagmes comprenant *mener* (c'est nous qui mettons en gras les informations pertinentes pour l'extraction des collocations)

Ces informations structurées peuvent ensuite être filtrées pour extraire une liste pertinente de collocations transdisciplinaires. Une première expérimentation en ce sens a été réalisée avec succès. Un important filtrage manuel doit néanmoins être effectué puisqu'aucune désambiguïsation automatique ne peut être effectuée – en tout cas facilement – à partir du dictionnaire. On relève ainsi un très grand nombre de cooccurrences non pertinentes comme *allouer à qqn un traitement* ou *méthode aratoire* qu'il faut bien entendu exclure de la liste désirée.

Le filtrage manuel permet néanmoins d'extraire assez rapidement un sous-ensemble d'éléments *a priori* pertinents pour notre projet. A titre d'exemple, nous listons dans le Tableau 1 un sous-ensemble de collocations *recherche(s)* -Adj extraites automatiquement et filtrées manuellement.

	recherche	appliquée
	recherche	bibliographique
	recherches	épistémologiques
	recherche	fondamentale
	recherche	inaccessible
	recherche	infructueuse
	recherche	prévisionnelle
	recherche	pure
	recherche	scientifique
	recherche	spéculative
	recherche	statistique
	recherche	stérile
	recherche	tâtonnante
	recherche	technique
	recherche	théorique
premières	recherches	
	recherches	récentes

Tableau 1 : Ensemble des collocations *recherche(s)* + Adjectif extraites automatiquement et filtrées manuellement

4. Extraction des collocations transdisciplinaires à partir de corpus, combinaison et évaluation des données extraites

Notre objectif final est de constituer des ressources lexicales utilisables. Nous souhaitons ainsi compléter les données extraites du *TLF* à l'aide de données extraites automatiquement de corpus d'écrits scientifiques, en utilisant l'analyseur syntaxique Syntex développé par Didier Bourigault (2007). Nous pensons que ces deux types de ressources sont complémentaires : les données du *TLFi* peuvent renvoyer à des collocations rares mais utiles (qui, du fait de leur faible fréquence seraient écartées lors d'une extraction automatique sur corpus exploitant des critères de fréquence et de répartition) alors que les données extraites des corpus correspondront à des données plus récentes, mais peut-être moins nombreuses. Dans le cadre de la méthode

automatique, les collocations du lexique transdisciplinaire sont obtenues à partir d'un corpus d'écrits scientifiques diversifiés de 2 millions de mots[7], auquel l'analyseur Syntex a été appliqué. Les noms retenus sont les noms les plus fréquents qui apparaissent à la fois dans les trois disciplines du corpus (économie, linguistique et médecine). Cette liste de noms est identique à celle qui a été exploitée pour l'extraction semi-automatique des collocations du *TLF*. Les collocations de type N-Adjectif et V-N sont simplement obtenues en extrayant les relations syntaxiques de dépendance de type épithète et objet direct apparaissant au moins trois fois dans deux des trois disciplines. Pour les relations de type V-N, les verbes *être* et *avoir* ont été exclus. Le Tableau 2 ci-dessous présente un extrait de collocations de type N-Adjectif obtenues automatiquement et dont on a enlevé les adjectifs qu'on peut considérer comme des mots grammaticaux (i.e. *autre, même, différent,* les adjectifs ordinaux et cardinaux).

Collocation N-Adjectif	**Fréquence**
étude récente	54
études empiriques	50
présente étude	31
approches théoriques	30
analyse statistique	24
analyse factorielle	20
études antérieures	18
étude précédente	15
étude préliminaire	15
argument supplémentaire	14
études complémentaires	13
analyse fine	12
analyse automatique	11
étude comparative	10
étude spécifique	9

Tableau 2 : Collocations transdisciplinaires de type N-Adj extraites automatiquement

Sur un premier test effectué sur 8 noms transdisciplinaires[8], les données extraites du *TLFi* semblent bien plus nombreuses que celles qui sont extraites du corpus. Elles comportent aussi beaucoup moins d'erreurs que les données automatiquement extraites de corpus où les erreurs d'analyse syntaxique introduisent un bruit non négligeable. Il apparaît en revanche difficile de porter un jugement clair sur certaines collocations extraites du *TLFi* (en tout cas, sans plus de contexte). Par exemple, l'expression *recherche infructueuse* qui relève clairement de la langue courante (Ex : *ma recherche d'appartement s'est révélée infructueuse*), apparaît-elle dans les écrits scientifiques ? A-t-elle sa place dans un lexique scientifique transdisciplinaire ?

Afin d'établir une liste de collocations de ce champ sémantique, nous proposons de constituer d'abord une liste de collocations candidates issues des deux ressources, corpus et *TLFi*, que nous avons exploitées. Y figureront en tête les collocations trouvées dans les deux ressources puis celles qui ne sont présentes que dans le *TLFi* puis celles qui ne sont présentes que dans le corpus (triées selon leur fréquence). Cette première liste sera évaluée par des linguistes experts du domaine. Notre hypothèse actuelle est que les collocations communes aux deux méthodes d'extraction seront les mieux évaluées ; l'analyse des résultats permettra de mieux cerner ce qu'apporte chaque ressource et, éventuellement, de se faire de nouvelles idées sur de meilleures stratégies pour les combiner.

Lors de l'évaluation, nous demanderons aussi aux linguistes d'établir si, selon eux, l'expression relève du lexique scientifique transdisciplinaire, s'ils l'emploieraient, s'ils la relèveraient pour une application d'aide à la rédaction et aussi, si l'expression relève de la langue générale ou de la terminologie du domaine.

[7] Le corpus comprend le corpus d'articles scientifiques KIAP de l'équipe de Kjersti Fløttum, augmenté par nos soins de rapports de recherche et thèses. Le corpus de 2 millions de mots se répartit équitablement entre les domaines de la médecine, de la linguistique et de l'économie.

[8] Les noms : *analyse, approche, argument, concept, démarche, étude, idée, recherche.*

Bibliographie

[Bourigault, 2007] Bourigault, D. (2007). *Syntex, analyseur syntaxique opérationnel*. Thèse d'Habilitation à Diriger des Recherches. Université Toulouse le Mirail, juin 2007.

[Dendien&Pierrel, 2003] Dendien, J. & Pierrel, J.-M. (2003). « Le Trésor de la Langue Française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence », *TAL*, vol. 44 - n°2, 11-39.

[Fontenelle, 1997] Fontenelle Th. (1997). *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. (Lexicographica /Series maior). Tübingen, Niemeyer Verlag.

[Hausmann, 1989] Hausmann F. J. (1989). Le dictionnaire de collocations. In Hausmann F.J., Reichmann O., Wiegand H.E., Zgusta L. (eds), *Wörterbücher : ein internationales Handbuch zur Lexicographie. Dictionaries. Dictionnaires*. Berlin/New-York, De Gruyter, 1010-1019.

[Hausmann, 1996] Hausmann F. (1996). La syntagmatique dans le *TLF* informatisé, in *Autour de l'informatisation du TLF*, Actes du Colloque International de Nancy (29-31 mai 1995), D. Piotrowski (ed.). Paris, Didier, 51-77.

[Henry, 1996] Henry F. (1996). Pour une informatisation du *TLF,* , in *Autour de l'informatisation du TLF*, Actes du Colloque International de Nancy (29-31 mai 1995), D. Piotrowski (ed.). Paris, Didier, 79-139.

[Mel'čuk, 1998] Mel'čuk I. (1998). Collocations and Lexical Functions. In A. P. Cowie (ed.), *Phraseology. Theory, Analysis and Applications*. Oxford, Clarendon Press, 23-53.