

Techniques de TAL et corpus pour faciliter les formulations en anglais scientifique écrit

Marie-Paule Jacques, Laura M. Hartwell, Achille Falaise

▶ **To cite this version:**

Marie-Paule Jacques, Laura M. Hartwell, Achille Falaise. Techniques de TAL et corpus pour faciliter les formulations en anglais scientifique écrit. 20e conférence sur le Traitement automatique des langues (TALN'2013), Jun 2013, Les Sables d'Olonne, France. hal-01911451

HAL Id: hal-01911451
https://hal.science/hal-01911451

Submitted on 2 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Techniques de TAL et corpus pour faciliter les formulations en anglais scientifique écrit

Marie-Paule Jacques[1] Laura Hartwell[1] Achille Falaise[2]
(1) Univ. Grenoble Alpes, UJF & LIDILEM, F-38040 Grenoble
(3) Univ. Grenoble Alpes, UPMF & LIG-GETALP, F-38040 Grenoble
(marie-paule.jacques,laura.hartwell)@ujf-grenoble.fr,
achille.falaise@imag.fr

RÉSUMÉ

Nous présentons l'adaptation de la base d'écrits scientifiques en ligne Scientext pour un « nouveau » public : chercheurs et autres auteurs français d'écrits scientifiques, ayant besoin de rédiger en anglais. Cette adaptation a consisté à ajouter dans la base des requêtes précodées qui permettent d'afficher les contextes dans lesquels les auteurs d'articles scientifiques en anglais expriment leur objectif de recherche et à enrichir l'interface ScienQuest de nouvelles fonctionnalités pour mémoriser et réafficher les contextes pertinents, pour faciliter la consultation par un public plus large. Les nombreuses descriptions linguistiques de la rhétorique des articles scientifiques insistent sur l'importance de la création et de l'occupation d'une « niche » de recherche. Chercheurs et doctorants ont ici un moyen d'en visualiser des exemples sans connaître sa formulation *a priori*, via nos requêtes. Notre évaluation sur le corpus de test en donne une précision globale de 86,5 %.

ABSTRACT

NLP and corpus techniques for finding formulations that facilitate scientific writing in English

This paper presents adaptations of the query options integrated into the online corpus Scientext so as to better serve a new audience: French scientists writing in English. We added pre-coded queries that display the contexts in which authors of scientific articles in English state their research objective. Furthermore, new functional options enrich the ScienQuest interface allowing results to be filtered for noise and then saved for consultation by a larger public.

Previous studies on the scientific discourse and rhetoric of scientific articles have highlighted the importance of establishing and occupying a research niche. Here, francophone researchers and doctoral students without prior discursive knowledge, can access authentic and multiple ways of formulating a research objective. Our evaluation of a test corpus showed an overall accuracy of 86.5 %.

MOTS-CLÉS : anglais, patrons lexico-syntaxiques, ScienQuest, Scientext.

KEYWORDS : ESP, lexico-syntactic patterns, ScienQuest, Scientext.

© ATALA

1 Introduction

Les chercheurs et jeunes chercheurs, quelle que soit leur discipline, sont amenés à produire des écrits scientifiques en anglais, ne serait-ce que pour le résumé de leurs articles. Si dans certaines disciplines, notamment les sciences dites « dures », la qualité de la langue n'est pas le premier attendu, il reste nécessaire de maîtriser les formulations par lesquelles se manifeste l'apport singulier de l'article, car elles en constituent souvent la « vitrine » qui aide les lecteurs potentiels à décider si l'article vaut le temps d'une lecture.

Cependant, les compétences des auteurs scientifiques sont sur ce point très variables et les chercheurs qui ne sont pas par ailleurs spécialistes de langue et de littérature n'ont pas nécessairement la disponibilité de se consacrer à un apprentissage formel, tout en ayant besoin de déterminer la formulation appropriée. Par exemple, les questions qui peuvent se poser lors de la rédaction concernent les expressions habituelles de l'objectif de la recherche, des hypothèses des chercheurs, des références aux travaux antérieurs, de l'apport spécifique de la recherche dans le champ disciplinaire, etc.

Nous décrivons ici l'adaptation à ce type de besoin d'une ressource en ligne, la base d'écrits scientifiques Scientext (Tutin *et al*, 2009 ; Falaise *et al.*, 2011a, Tutin *et al.,* à paraître*)*. Nous avons utilisé les outils et techniques du TAL et de la linguistique de corpus pour enrichir la base de requêtes avancées, pré-codées et mises à disposition pour la sélection de contextes sur des bases *rhétorico-sémantiques*. La double particularité de ces requêtes est que d'une part elles sont fondées sur les différents travaux relatifs à l'organisation rhétorique du texte scientifique pour proposer une entrée « par la fonction » et non « par la forme », d'autre part elles tirent parti de la puissance d'expression autorisée par l'annotation des textes en relations syntaxiques, ce qui permet de s'affranchir de variations liées à la linéarité et d'exprimer des dépendances entre les différents éléments constituant la cible visée.

L'intérêt d'une recherche *via* la fonction rhétorico-sémantique est de permettre au chercheur d'ignorer *a priori* le vocabulaire à employer et de découvrir les constructions possibles à travers les exemples authentiques auxquels une requête lui permet d'accéder. Par exemple, l'objectif de recherche s'énonce aussi bien sous la forme *"Here, we investigate the evolution of one of the most striking examples of sexual conflict in hermaphrodites..."* que "*The aim of this study was to examine the effect of ocean climate on foraging success in this deep-diving marine mammal ..."*. Nous nous focalisons ici plus particulièrement sur la formulation de l'objectif de recherche, à soigner à la fois parce qu'elle est généralement doublement présente, dans les résumés et dans l'introduction de l'article, et parce qu'elle signale la spécificité de la recherche qui fait l'objet de l'article dans le champ disciplinaire.

Notre article se structure ainsi : dans un premier temps, nous exposons les travaux sur la rhétorique de l'article scientifique et la façon dont ils ont déjà été exploités dans le cadre du TAL. Puis nous explicitons les besoins auxquels

 © ATALA

notre ressource en ligne veut répondre. Enfin nous décrivons et évaluons notre travail d'enrichissement de Scientext.

2 L'article scientifique et ses fonctions rhétoriques

Depuis l'étude empirique de Sinclair, Jones et Daley dans les années 1970 (2004), puis les travaux de Swales (1990/2004), les caractéristiques du genre *article scientifique* ont été explorées selon des axes divers : variations lexicales entre articles de recherche et autres types d'écrits scientifiques (Poudat et Follette, 2012), spécificités lexicales de certaines sections – par exemple l'emploi de certains verbes dans les résumés (Hartwell, 2013, Hartwell et Jacques 2012) ou la saillance d'items lexicaux (Gledhill, 2000) –, évolution des « patterns » lexico-grammaticaux selon les sections dans des articles de biomédecine (Saber, 2012), ou encore positionnement et auto-représentation (Hyland, 2004 ; Hyland, 2012), « voix » de l'auteur selon la discipline (Fløttum, Kinn et Dahl, 2006)... Loin d'avoir inventorié la totalité des travaux consacrés à l'écrit scientifique, particulièrement en « anglais de spécialité », les références qui précèdent ne donnent qu'un aperçu de leur foisonnement. Celui-ci tient sans doute au fait que, pour un scientifique, la maitrise du *discours* et des *genres* associés (Rastier, 2001) participe de la compétence de chercheur. En outre, si l'on en croit Pontille (2007), dans les sciences expérimentales, la normalisation de la structure de l'article de recherche a accompagné la fixation de la démarche scientifique à la fois comme production de connaissances et comme pratique sociale.

L'article scientifique réalise donc les conventions à la fois discursives et pratiques liées au fait même de faire de la science, ce qui est à la fois une contrainte et un cadre structurant. En effet, cet aspect conventionnel se traduit par des fonctions rhétoriques régulières dont on suppose qu'elles se formulent par des expressions linguistiques récurrentes. Cette hypothèse sous-tend notamment les travaux de Teufel sur ce qu'elle appelle *Argumentative Zoning*, à partir desquels ont notamment été élaborés des systèmes de résumé automatique (Teufel, 1998 ; Teufel *et al.*, 1999) et d'attribution automatique de citations (Siddharthan *et al.*, 2007 ; Teufel *et al.*, 2006). Ces recherches s'appuient sur la systématicité, et même plus, sur la nécessité des mouvements argumentatifs étudiés. Il serait en effet inenvisageable de faire de la recherche sans citer d'autres auteurs et/ou se positionner par rapport aux travaux antérieurs.

Une autre opération nécessaire de l'article de recherche est la création d'un espace de recherche (*creating a research space*), qui implique trois étapes (*moves*) : 1. établir un territoire de recherche ; 2. établir une *niche* dans ce territoire ; 3. occuper cette niche (Swales et Feak, 2004). Pour cette dernière, la mention de l'objectif ou des aspects majeurs de l'étude est obligatoire : *« outlining purposes or stating the nature of the present research »* (p. 244). C'est par cette mention que l'auteur occupe cet espace de recherche. Sa formulation représente de ce fait une compétence nécessaire pour les chercheurs.

 © ATALA

3 Besoins des chercheurs et ressources en ligne

La création de l'espace de recherche s'opère majoritairement dans l'introduction de l'article et même dès le résumé de l'article. Or, même les revues francophones demandent, pour la plupart, un résumé en anglais, ce qui implique que, potentiellement, tout chercheur français est confronté à la formulation en anglais de points aussi cruciaux que la singularité et l'apport de sa recherche.

Carter-Thomas *et al.* (à paraitre) ont mis en évidence que les nuances, subtilités et spécificités d'expression en anglais sont malaisées à maitriser pour les chercheurs français, même chevronnés, même habitués à rédiger en anglais. Les approximations langagières peuvent être pénalisantes, y compris dans les disciplines pour lesquelles la correction linguistique n'est pas un enjeu majeur (les sciences dites « dures », notamment). Mais l'emploi du temps des chercheurs et encore plus des apprentis-chercheurs (doctorants) ne leur laisse pas le loisir de se consacrer à un apprentissage poussé de la langue. Il existe divers ouvrages de conseils pour la rédaction (par exemple Matthews et Matthews, 2008 ; Swales et Feak, 2004), cependant, aussi bien faits qu'ils soient, ils sont rarement opératoires et ne permettent généralement pas de répondre rapidement et « en direct » à la question « cette formulation est-elle appropriée ? » ou « comment exprime-t-on... ? ».

Les exemples authentiques d'articles de recherche publiés seraient à même de fournir une aide opératoire en ce qu'ils permettent de voir concrètement « ce qui se dit » dans telle ou telle discipline. Dans le champ de l'anglais de spécialité, les mérites de l'accès aux corpus d'écrits authentiques sont de plus en plus reconnus (Boulton *et al.*, 2006). La maison d'édition Springer, à travers son site Springer Exemplar[1] offre un tel accès en permettant des concordances sur des milliers d'articles publiés. Mais on ne peut sélectionner les contextes que par une requête sur un mot ou une expression, ce qui est très utile si l'on veut vérifier l'emploi de tel verbe, tel nom ou tel adjectif, mais ne permet pas au scripteur hésitant de déterminer avec quelle forme ou expression traduire le mouvement rhétorique requis.

C'est pourquoi nous proposons de réutiliser la base en ligne d'écrits scientifiques Scientext[2], originellement conçue pour la recherche linguistique, en l'adaptant et l'enrichissant pour offrir, à côté des concordances classiques sur la forme, le lemme, la catégorie, un accès par la signification. Il s'agit bien de permettre un accès direct à des contextes dont nous avons vérifié qu'ils remplissent telle ou telle fonction rhétorique (ici l'expression de l'objectif de recherche). De telles requêtes existent déjà pour le français, mais elles concernent l'étude du positionnement de l'auteur et n'ont pas été complètement reproduites pour l'anglais. L'offre de requêtes précodées en anglais est jusqu'ici limitée aux citations.

[1]http://www.springerexemplar.com/ Consulté le 12/04/2013.
[2]Disponible à l'adresse : http://scientext.msh-alpes.fr Consulté le 12/04/2013.

© ATALA

4 Adaptation de Scientext

La base Scientext a été créée à destination des linguistes, pour permettre l'étude des traits du discours scientifique. L'écrit scientifique est représenté à travers deux corpus :

- en français, des thèses, articles de conférences et articles de revues essentiellement dans des disciplines de SHS (4,8 millions de mots) ;

- en anglais, des articles scientifiques de biologie et médecine (14,8 millions de mots).

Les textes mis à disposition dans la base ont ceci de remarquable par rapport à d'autres bases textuelles qu'ils ont été analysés syntaxiquement. La sélection d'un corpus et son interrogation se font à travers l'interface ScienQuest, mise au point par A. Falaise (Falaise *et al.*, 2011b ; 2012). Celle-ci ajoute aux fonctionnalités classiques de concordances sur formes, lemmes et catégories la possibilité de spécifier des contraintes de relations syntaxiques entre les éléments de la requête.

Nos requêtes ont été construites sur un corpus d'étude plus restreint et testées sur le reste de la base, selon la démarche exposée dans cette section.

4.1 Recueil des données : corpus et méthode

Suivant une démarche courante en linguistique de corpus et TAL, nous avons dans un premier temps constitué une liste des phrases indiquant l'objectif de la recherche, afin d'en observer les différentes formulations et d'avoir une liste de référence pour l'évaluation initiale. Cette liste a été dressée à partir d'un sous-ensemble de 600 articles choisis selon deux critères, la variété des revues de publication et l'origine anglophone de l'article, appréhendée par l'adresse de l'auteur : nous avons retenu des articles provenant d'universités d'Angleterre, du Canada et des États-Unis d'Amérique. Cette restriction n'est pas une garantie absolue que l'auteur soit un anglophone natif, mais elle limite tout de même la quantité de non-natifs. Cette sélection de 600 articles a été opérée de façon automatique, par un script extérieur à ScienQuest.

Pour la constitution de la liste de phrases, nous nous sommes limités aux résumés. Le recueil a été effectué en partie manuellement, donc se concentrer sur les résumés a accéléré la tâche. De plus, dans la structuration d'un article, cette étape de définition et occupation d'une niche est préalable et intervient dans le résumé et/ou dans l'introduction.

Une centaine des 600 résumés n'indiquait pas de façon claire et délimitée l'objectif de recherche, celui-ci était en quelque sorte dilué dans l'exposé de ce que la recherche couvrait. Il était accessible par inférence, mais laissé implicite. Cent autres résumés n'ont pas contribué à la modélisation parce que les formulations employées soit évoquaient un travail sur l'hypothèse de recherche (*To test the hypothesis...*), soit employaient une langue si peu « standard » que le doute était possible sur la langue native de l'auteur, soit

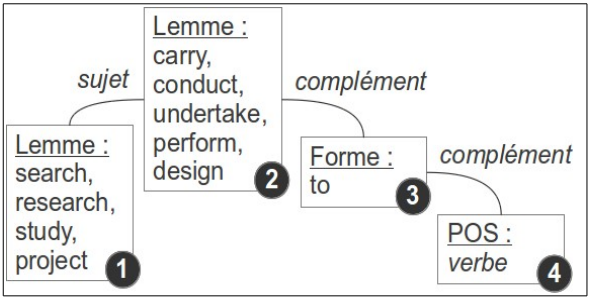
 © ATALA

ne présentaient pas la régularité suffisante pour une modélisation, c'est-à-dire qu'il eut fallu presque une requête différente par phrase. La modélisation et l'élaboration de requêtes se sont donc appuyées sur 400 énoncés.

4.2 Modéliser l'expression des objectifs de recherche

La modélisation des expressions et structures employées par ces énoncés pour stipuler l'objectif de la recherche s'apparente à la construction de patrons lexico-syntaxiques (Condamines *et al.*, 2006 ; Jacques *et al.*, 2006). Il s'agit de repérer à la fois les régularités sous-jacentes et les caractéristiques discriminantes des énoncés-cibles. Il faut en établir une grammaire qui cerne les traits par lesquels obtenir la sélection des énoncés-cibles et seulement ceux-là. La modélisation et la construction des requêtes pour ScienQuest procèdent donc par 1. typification des occurrences, 2. repérage et élimination des traits non discriminants. Ce dernier temps réside dans l'adaptation successive des patrons, telle que montrée dans (Rebeyrolle *et al.*, 2001).

C'est à cette étape que l'annotation syntaxique et le supplément potentiel de contraintes qu'elle comporte montre toute sa puissance. Par exemple, une requête (figure 2, plus loin) qui modélise le patron détaillé dans la figure 1 peut « capter » aussi bien l'exemple 12 de la section 5.1 que :



(1) A descriptive ***study❶*** *of adult cystic fibrosis patients [...] was* ***conducted❷*** ***to❸*** ***evaluate❹*** *the prevalence of osteoporosis*

(2) ***To❸*** ***determine❹*** *the frequency, risk factor and mortality of nosocomial pneumonia a prospective* ***study❶*** *was* ***conducted❷***

Figure 1 – Un exemple de patron.

Comme on le voit avec (1), le recours aux relations syntaxiques permet de ne pas se soucier spécifiquement des cas d'insertion de constituants (adjectifs, compléments de noms, adverbes...), alors que dans le cas de patrons bâtis uniquement sur de l'étiquetage morphosyntaxique, il est nécessaire de prévoir à l'avance une distance possible entre les éléments.

De même, (2) montre que l'on n'a pas à spécifier l'ordre des constituants : si dans le patron est inscrite une relation entre un verbe et un complément ou entre un verbe et un sujet, ceux-ci peuvent aussi bien se trouver avant ou après, en début ou en fin de phrase, à proximité ou à distance.

Tirant parti de cette puissance, une douzaine de patrons différents a été élaborée pour prendre en compte la variété des expressions.

© ATALA

5 Expression de l'objectif de recherche en anglais

5.1 Analyse linguistique

Plusieurs indices sont discriminants pour une identification automatique de l'expression de l'objectif de recherche. En premier lieu, on peut remarquer que c'est l'occasion de voir apparaître le chercheur lui-même, sous la forme d'un pronom personnel tel que *I* ou *we*. Toutefois, cet indice ne constitue pas à lui seul un marqueur probant, nous l'avons donc associé à d'autres éléments linguistiques : certains verbes précis, ainsi que le déictique *here*, qui, positionné en début de phrase, restreint la portée de la proposition au contexte immédiat, c'est-à-dire l'article.

En second lieu, les termes mêmes signifiant *objectif*, à savoir *purpose, goal, aim, objective*, combinés au verbe *be* et un complément introduit par *to*, pour signifier un but, sont de bons marqueurs des contextes recherchés.

Un tel complément introduit par *to* ou *in order to* apparaît régulièrement dans les contextes recueillis, mais aussi dans d'autres contextes, comme la citation d'autres travaux. Pour limiter son ambiguïté, nous avons contraint sa position à la tête de phrase ou nous l'avons associé à d'autres contraintes lexico-syntaxiques.

La déixis, quand elle permet de désigner l'article lui-même ou l'étude qui y est exposée, fournit un bon indice à travers des constructions comme *(in) this study*, *(in) the present study*.

Enfin, certains verbes tels que *present, determine, assess, describe,* en ce qu'ils sont relativement spécialisés dans l'expression de la recherche, sont des marqueurs satisfaisants, à condition de limiter leur occurrence aux contextes dans lesquels leur sujet est le pronom *we* ou *I*, désignant le ou les auteurs eux-mêmes.

Mais cette restriction peut s'avérer insuffisante. De manière générale, une difficulté récurrente tient à l'ambiguïté entre l'expression de l'objectif et l'expression de la méthode. Un verbe tel que *compare* est à cet égard exemplaire. Dans *« We compare Monte Carlo Markov chain analysis of two very different measures of hypertension in the simulated Genetic Analysis Workshop 13 data to examine how choice of measure affects the results. »,* les chercheurs indiquent ce à quoi ils veulent aboutir mais *compare* permet d'expliquer ce que les chercheurs font et non *stricto sensu* la « niche » occupée. La moitié de ses occurrences ne correspond pas aux contextes recherchés, nous l'avons donc finalement éliminé des diverses listes de verbes entrant dans les requêtes.

Ce repérage d'indices a abouti à l'élaboration de douze patrons, pour lesquels voici des exemples de contextes visés (en caractère gras les éléments retenus pour la construction des requêtes) :

1. ***Here**, **we report** the results of a survey to assess the prevalence of drg in*

© ATALA

a globally representative panel of disease-associated meningococci.

2. ***Here**, a statistical **test** to detect gene conversion [...] **is presented**.*
3. ***In order to determine** which foods might be related to disease activity in UC a new method of dietary analysis was developed and applied.*
4. ***To understand** the physiological processes responsible for elevated Cd accumulation in shoots [...], Cd uptake and translocation were studied [...]*
5. ***In this** retrospective **review**, we examine whether progression to ESRF can be predicted and whether treatment [...]*
6. ***In the current study** we report the isolation and preliminary characterization of homologous proteins from goat seminal plasma.*
7. ***This article outlines** the evolution of a community pharmacy-based supervised consumption of methadone program in Grater Glasgow.*
8. ***The present study addresses** the relationship of protein folding propensities to the evolutionary relationship between residues.*
9. ***Our aim was to determine** the effects of taking a red clover-derived isoflavone supplement daily for 1 year on mammographic breast density.*
10. ***We present** Homology Induction (HI), a new approach to inferring homology.*
11. ***We aimed to assess** warfarin treatment in primary health care*
12. ***This study** was **undertaken to characterize** the expression of chemokine receptors*

Pour aperçu des requêtes dans ScienQuest[3], voici celle qui correspond à (12) :

```
// Un verbe d'« étude » ayant pour sujet un nom désignant la recherche et pour complément un SP formé de to et un verbe quelconque
$research=search,research,study,project          // liste noms de recherche
$verb4=carry,conduct,undertake,perform,design    // liste verbes d'étude
Main = <form=$research,#2> && <lemma=$verb4,#1> && <lemma=to,#4> && <cat=/V/,#5> :: ((SUJ,#1,#2) OR (SUJCOMP,#2,#1)) AND (PREP,#1,#4) AND (NOMPREP,#4,#5) ;    // règle principale (SUJ, SUJCOMP, PREP, NOMPREP désignent les relations syntaxiques entre les unités lexicales ou grammaticales)
```

FIGURE 2 – Un exemple de requête dans ScienQuest.

5.2 Evaluation

Comme souvent lorsqu'il s'agit de définir les « marqueurs » linguistiques d'un certain contenu sémantique, la projection sur les corpus des requêtes élaborées « ramène » aussi bien les contextes visés que d'autres contextes qui n'ont pas la signification souhaitée – notamment en raison de la forte

[3]L'ensemble des requêtes est rendu disponible sur le site de Scientext.

© ATALA

ambiguïté entre objectif de recherche et méthode, soulignée plus haut.

Nous avons donc évalué chacune de nos requêtes : dans un premier temps sur le corpus d'étude de 600 articles, dans un deuxième temps sur le reste du corpus anglais dans Scientext, soit environ 8000 articles. Nous avons limité la mesure sur ce dernier à 500 contextes mais certaines requêtes n'en fournissent pas autant.

Le tableau 1 récapitule les différentes requêtes, le nombre de contextes renvoyé par chacun et la précision pour chaque corpus.

Patrons	Corpus d'étude		Corpus de test	
	Nombre d'occ.	Précision	Nombre d'occ.	Précision
1 (Here, we...)	26	88,5 %	342	87,1 %
2 (Here, a N is...)	3	100 %	22	59,1 %
3 (In order to...)	12	66,7 %	104	86,5 %
4 (To V...)	49	34,7 %	393	77,1 %
5 (In this study)	34	88,2 %	421	88,6 %
6 (In the present study)	4	80 %	128	89,1 %
7 (This paper V)	94	81,9 %	498	80,5 %
8 (The present paper V)	12	75 %	156	87,8 %
9 (Our aim is to)	63	90,5 %	496	98,8 %
10 (We present)	154	77,9 %	500	87 %
11 (we V to)	213	63,8 %	259	82,6 %
12 (this study V to)	13	84,6 %	192	87,5 %

TABLE 1 – Mesures de précision des requêtes pour l'expression de l'objectif de recherche en article scientifique en anglais

Nous n'avons pris en compte que la précision car la tâche définie ici n'implique pas de traquer tous les résultats possibles pour une requête. Ce qui est important, c'est que les requêtes soient formulées de telle sorte qu'elles

© ATALA

fournissent au chercheur-scripteur une idée satisfaisante de la façon dont « ça peut s'écrire », c'est-à-dire qu'elles couvrent les diverses variantes d'un même type de formulation. Par exemple, les 40 premiers résultats des requêtes 7 et 8 ci-dessus montrent 34 combinaisons de verbes et sujets différentes, ce qui semble suffisamment informatif pour un temps de consultation restreint.

Par ailleurs, les résultats surprenants de la ligne 4 réclament explication. Dans notre corpus d'étude, les résultats ont en fait été « pollués » par des verbes indiquant la méthode plus que l'objectif (par ex. *To overcome this limitation*...), toutefois ces verbes étaient proportionnellement moins présents dans le corpus de test : "*to overcome*" par ex. apparaît 2 fois sur 49 dans le corpus d'étude et 4 fois sur 393 dans le corpus de test.

L'ensemble des requêtes sur le corpus de test donne une précision globale moyenne de 86,5 % : 3512 contextes vérifiés (par une locutrice native d'anglais), 475 non validés. Certaines occurrences correspondent à plusieurs requêtes, sur le corpus d'étude, le recouvrement était d'environ 9 %.

Avoir des patrons qui donnent des résultats les moins bruités possible ne suffit pas si on ne s'adresse ni à des linguistes ni à des étudiants de langue mais à un public qui n'a pas forcément la compétence pour faire lui-même la distinction entre énoncés pertinents et bruit. La dernière adaptation pour atteindre notre objectif concerne l'interface de ScienQuest elle-même.

6 Mémoriser les contextes pertinents

A l'origine, Scientext et son interface ScienQuest ont été prévus pour offrir à l'utilisateur 3 niveaux de requêtes : un niveau de concordancier dans lequel l'utilisateur est guidé pour fabriquer des requêtes pouvant combiner formes, lemmes, catégories morpho-syntaxiques et relations syntaxiques ; un niveau de requêtes sémantiques dans lequel l'utilisateur sélectionne une requête sémantique pré-codée et enfin un niveau avancé dans lequel l'utilisateur code lui-même sa requête avec le langage d'interrogation sous-jacent (Falaise *et al.*, 2011b). Mais, quel que soit le mode d'interrogation choisi, on obtient des résultats bruts et par conséquent bruités.

La modification actuelle consiste donc à pouvoir mémoriser (à terme, sur le serveur) et recharger autant que de besoin une série de contextes validés. La définition ici exposée de requêtes sémantiques pour atteindre dans les textes la formulation des objectifs de recherche ne se limitera donc pas à la mise à disposition des requêtes élaborées mais aussi des contextes débarrassés de leurs réponses non pertinentes.

La figure 3 montre une partie de cette nouvelle fonctionnalité : à la fin d'une liste de contextes – dont certains invalidés (dé-sélection) –, une commande pour sauvegarder soit les résultats seuls à exploiter dans un autre logiciel, soit la sélection à recharger ultérieurement dans ScienQuest.

© ATALA

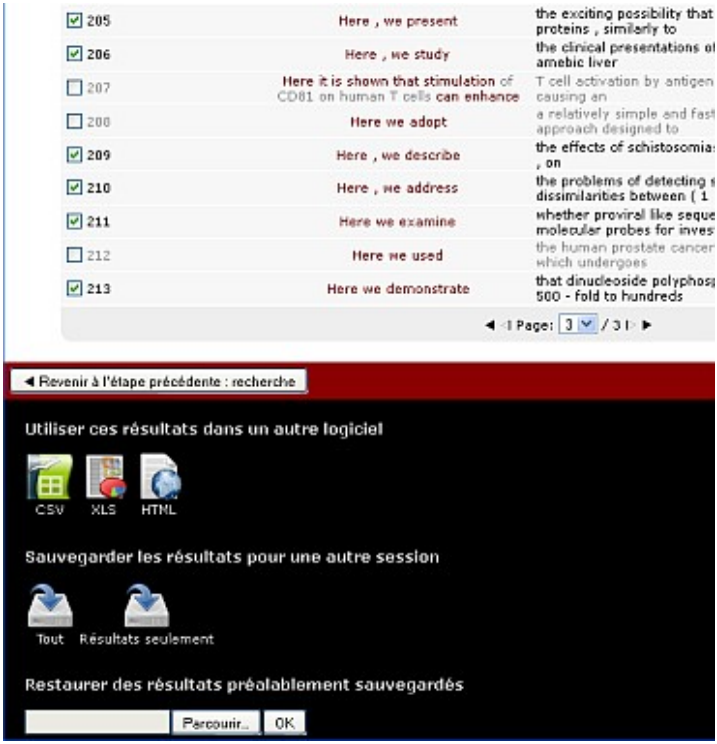


FIGURE 3 – Validation des résultats et sauvegarde dans ScienQuest.



FIGURE 4 – Chargement d'une requête et ses résultats dans ScienQuest.

© ATALA

Lors d'une autre session, au lieu de faire une nouvelle recherche, il sera possible de recharger la requête et les résultats vérifiés (Figure 4).

A terme, l'utilisateur disposera directement sur le site de Scientext non seulement des requêtes, à utiliser en bloc ou une par une, mais aussi d'un certain nombre de contextes validés, ce qui constitue une fonctionnalité utile aussi pour l'enseignement-apprentissage des langues.

7 Conclusion

L'adaptation de la base Scientext en vue d'offrir un accès immédiat à des contextes pertinents pour l'expression de l'objectif de recherche en anglais a impliqué trois étapes :

1. le recueil en corpus d'un échantillon d'occurrences des énoncés visés ;

2. la modélisation de ces énoncés sous la forme de patrons lexico-syntaxiques traduits en requêtes dans ScienQuest ;

3. dans l'interface ScienQuest, la validation et la mémorisation des énoncés pertinents pour limiter les affichages ultérieurs à ces seuls énoncés.

Ce travail vise à combler un manque lié au fait que la plupart des outils d'interrogation de gros corpus offre des possibilités de concordances ou d'extractions à partir des lemmes, formes ou catégories morphosyntaxiques (pour certains) mais pas à partir des significations elles-mêmes. Or, une difficulté fréquente pour un locuteur non-natif qui doit rédiger dans une langue seconde réside dans sa méconnaissance des formes mobilisées pour un certain sens. Les corpus qui forment la base Scientext peuvent ainsi être utilisés pour visualiser la diversité des formulations pour une même intention communicative. Nos requêtes s'ajoutent à celles qui existent déjà en anglais et en français pour la formulation des citations d'autres travaux, des hypothèses de recherche, du positionnement de l'auteur, élaborées selon une démarche similaire à celle que nous présentons.

L'adaptation de l'interface concerne l'ensemble de la base :corpus anglais et français. Pour mesurer l'utilité des nouvelles fonctionnalités, il est prévu de tester cette approche sur le corpus français dans le cadre de l'enseignement assisté par ordinateur, avec un public de non-francophones[4].

Un autre intérêt non négligeable du travail présenté ici est la réutilisation d'une ressource publique et ouverte à tous. Ce type de ressource étant relativement coûteux à constituer, pour rentabiliser les efforts et l'argent investis et pérenniser la ressource, il nous semble nécessaire de diversifier ses utilisations et de la rendre disponible pour de nouveaux publics et de nouveaux besoins. Cela passe, comme nous l'avons montré, par une évolution des fonctionnalités et par une augmentation continue des ressources connexes telles que les requêtes précodées. La collaboration du TAL et de la linguistique de corpus est ainsi optimale.

[4] Dans un projet impliquant A. Falaise, H. Tran et A. Tutin, du laboratoire LIDILEM.

© ATALA

8 Références

BOULTON, A. et WILHELM, S. (2006). Habeant Corpus-they should have the body. Tools learners have the right to use. *ASp,* 49-50, pages 155-170.

CARTER-THOMAS, S. & ROWLEY-JOLIVET, E. (à paraitre.) Rapporter la voix de l'autre dans les articles de recherche en anglais : problèmes et enjeux pour le chercheur francophone. *Le discours rapporté et ses marques : perspectives théoriques et didactiques*. Editions Aracne.

CONDAMINES, A. et JACQUES, M.-P. (2006). Le repérage de l'hyperonymie par un faisceau d'indices : mise en question de la notion de « marqueur ». *Journée "Textes et connaissances", Semaine de la Connaissance*, pages 185-194.

FALAISE, A., TUTIN, A., KRAIF, O., ROUQUET, D. (2012). ScienQuest: a treebank exploitation tool for non NLP-specialists. *Actes de COLING 2012*, Mumbai, Inde.

FALAISE, A., TUTIN, A. et KRAIF, O. (2011a). Exploitation d'un corpus arboré pour non spécialistes par des requêtes guidées et des requêtes sémantiques. in *Actes de TALN 2011 (Traitement automatique des langues naturelles)* Montpellier, pages 187-215.

FALAISE, A., TUTIN, A. et KRAIF, O. (2011b). Définition et conception d'une interface pour l'exploitation de corpus arborés pour non-informaticiens : la plateforme ScienQuest du projet Scientext. *TAL* 52(3), pages 103-128.

FLØTTUM, K., KINN, T. et DAHL, T. (2006). "We now report on ..." versus "Let us now see how ...": Author roles and interaction with readers in research articles. *In Academic Discourse across Disciplines* (Hyland et Bondi, 2006), pages 203-224.

HARTWELL, L. (2013). Corpus-informed descriptions: English verbs and their collocates in science abstracts. *In Etudes en didactique des langues* (Décuré, 2013), pages 79-95.

HARTWELL, L. & JACQUES, M.-P. (2012). A corpus-informed text-reconstruction resources for learning about the language of scientific abstracts. *in Actes de EuroCALL 2012*, Suède: pages 117-123.

HYLAND, K. (2004). Patterns of engagement: dialogic features and L2 undergraduate writing. *In Analysing Academic Writing: Contextualised Frameworks* (Ravelli et Ellis, 2004), pages 5-23.

HYLAND, K. (2012). Disciplinary Identities: Individuality and community in academic discourse. Cambridge: Cambridge University Press.

JACQUES, M.-P. et AUSSENAC-GILLES, N. (2006). Variabilité des performances des outils de TAL et genre textuel. Le cas des patrons lexico-syntaxiques. *TAL* 47(1), pages 11-32.

MANIEZ, F. (2012). A corpus-based study of adjectival *vs.* nominal modification in medical English . *In Corpus-informed research and learning in ESP: Issues and applications* (Boulton *et al*, 2012), pages 83-102.

© ATALA

MATTHEWS, J. R. et MATTHEWS, R. W. (2012). *Successful scientific writing: A step-by-step guide for the biological and medical sciences* (3*rd edition*). Cambridge: Cambridge University Press.

PONTILLE, D. (2007). Matérialité des écrits scientifiques et travail de frontières : le cas du format IMRAD. *in Sciences et frontières* (Hert et Paul-Cavallier, 2007), Fernelmont, pages 229-253.

POUDAT, C, et FOLLETTE, P. (2012). Corpora and academic writing: A contrastive analysis of research articles in biology and linguistics. *In Corpus-informed research and learning in ESP: Issues and applications* (Boulton *et al*, 2012), pages 167-192.

RASTIER, F. (2001). Eléments de théorie des genres. *Texto !* juin 2001 [en ligne]. <http://www.revue-texto.net/Inedits/Rastier/Rastier_Elements.html> (Consulté le 12/04/2013).

REBEYROLLE, J. et TANGUY, L. (2001). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de Grammaire* 25, pages 153-174.

SABER, A. (2012). Phraseological patterns in a large corpus of biomedical articles. *In Corpus-informed research and learning in ESP: Issues and applications* (Boulton *et al*, 2012), pages 45-82.

SIDDHARTHAN, A. et TEUFEL, S. (2007). Whose Idea Was This, and Why Does it Matter? Attributing Scientific Work to Citations. *HLT-NAACL,* pages 316-323.

SINCLAIR, J., JONES, S. DALEY, R. (2004). English collocation studies: The OSTI report. Krishnamurthy (Ed.), London : Continuum.

SWALES, J. M. (1990/2004). Genre Analysis: English in Academic and Research Settings. Cambridge : Cambridge University Press.

SWALES, J. M. AND FEAK, C. B. (2004). Academic writing for graduate students: Essential tasks and skills, Second edition. Ann Arbor (MI): U. of Michigan Press.

TEUFEL, S. (1998). Meta-discourse markers and problem-structuring in scientific articles. *Workshop on Discourse Structure and Discourse Markers*, ACL 1998, Montreal.

TEUFEL, S., CARLETTA, J. et MOENS, M. (1999). An annotation scheme for discourse-level argumentation in research articles. *EACL*. pages 110-117.

TEUFEL, S., SIDDHARTHAN, A. et TIDHAR, D. (2006). Automatic classification of citation function. *EMNLP,* pages 103-110.

TUTIN A., GROSSMANN F., FALAISE A., KRAIF O. (2009). Autour du projet Scientext : étude des marques linguistiques du positionnement de l'auteur dans les écrits scientifiques. *Journées Linguistique de Corpus*. Lorient.

TUTIN, A. et GROSSMANN, F., éditeurs (à paraitre). Autour du corpus Scientext : de la constitution d'un corpus d'écrits scientifiques à l'étude des marques du positionnement et du raisonnement. Presses Universitaires de Rennes.

© ATALA