Autour du projet Scientext : étude des marques linguistiques du positionnement de l'auteur dans les écrits scientifiques

Agnès Tutin, Francis Grossmann, Achille Falaise, Olivier Kraif

Autour du projet Scientext : étude des marques linguistiques du positionnement de l'auteur dans les écrits scientifiques

Agnès Tutin[*], Francis Grossmann[*], Achille Falaise[+], Olivier Kraif[*]
[*]Lidilem – Université Grenoble 3 –Stendhal
{agnes.tutin,francis.grossmann,olivier.kraif}@u-grenoble3.fr
[+]GETALP – LIG – Université Grenoble 1 – Joseph Fourier
achille.falaise@imag.fr

Résumé

Cet article présente le projet Scientext, qui a permis de constituer un corpus d'écrits scientifiques variés et des outils logiciels permettant d'effectuer une étude linguistique du positionnement et du raisonnement dans les écrits scientifiques, à travers des marques linguistiques. Nous retraçons ici les ressources développées pour le français dans le cadre du projet et présentons une étude de cas du positionnement de l'auteur, à travers l'étude des verbes de positionnement associés à un sujet auteur en sciences humaines.

Introduction

Le projet ANR Scientext[1] s'inscrit dans le domaine de la linguistique de corpus. Il poursuit deux objectifs : d'une part, constituer un corpus d'écrits scientifiques variés d'écrits scientifiques, permettant à la fois de comparer des ensembles de disciplines et des sous-genres scientifiques, comme la thèse ou l'article de recherche ; d'autre part, l'étude linguistique des marques explicites du positionnement et du raisonnement de l'auteur dans ce genre.

Nous décrivons ici dans un premier temps les objectifs du projet, les corpus constitués à cette fin et les annotations effectuées. Dans un second temps, nous abordons une problématique linguistique traitée dans le cadre du projet, l'étude des verbes de positionnement explicite en sciences humaines (linguistique, psychologie, sciences de l'éducation). Notre objectif est ici d'observer dans quelle mesure la variable disciplinaire apparaît déterminante dans l'emploi de la première personne et de ces marques de positionnement explicites, et parallèlement d'observer comment se construit la visibilité de l'auteur scientifique. Nous présentons enfin les modes d'exploitation du corpus mis en œuvre dans le site en ligne où, à la façon de Frantext, les corpus sont librement interrogeables par divers types de requêtes ("simples", "sémantiques", "avancées").

[1] Projet dans le cadre de l'ANR « Corpus et outils de la recherche en sciences humaines et sociales » (2007-2010). Site du projet : http://scientext.msh-alpes.fr.

1. Présentation du projet Scientext

1.1 Objectifs du projet

Le projet Scientext vise à produire des données et des analyses linguistiques sur les écrits scientifiques intéressant les linguistes, mais aussi les spécialistes de l'extraction d'information qui cherchent à identifier des passages spécifiques, comme les références à autrui (par exemple, Siddhartan & Teufel, 2007), véritable enjeu dans la veille technologique. A plus long terme, le projet pourrait également permettre la construction de bases de données lexicales, d'outils d'aide à la rédaction basés sur corpus (Kraif & Tutin 2009), ainsi que l'élaboration d'outils pour l'extraction d'informations scientifiques.

Scientext s'inscrit pleinement dans la linguistique de corpus, puisqu'il cherche à mettre en évidence les spécificités lexicologiques et énonciatives du genre des écrits scientifiques en se basant sur un ensemble de textes authentiques. Le projet recourt en outre aux techniques du traitement automatique des langues, à la fois pour l'analyse syntaxique du corpus (effectué à l'aide de l'analyseur de dépendance Syntex développé par Didier Bourigault (2007)), et pour l'interrogation du corpus qui exploite des requêtes complexes (Cf. Kraif 2008 ; Falaise et Tutin 2009), comme on le verra dans la troisième section. L'analyse du discours, la lexicologie et l'approche énonciative sont aussi convoquées pour l'analyse linguistique du positionnement et du raisonnement.

Trois équipes ont été impliquées dans ce projet. Le LIDILEM[2] (Laboratoire de Linguistique et de Didactique du Française Langue Etrangère et Maternelle, Université Grenoble 3-Stendhal) qui a coordonné le projet, a constitué le corpus d'écrits scientifiques français et l'interface informatique permettant d'exploiter le corpus. Le laboratoire Littérature Langage Société[3], de l'Université de Chambéry, a élaboré un corpus d'anglais académique de locuteurs non natifs. L'équipe LiCorn (Linguistique de Corpus, Université de Bretagne Sud[4]), a recueilli et traité un corpus d'anglais scientifique, principalement en sciences de la vie et médecine.

Le produit final réalisé est un site Web (adresse : http://scientext.msh-alpes.fr), à la façon de Frantext, permettant de sélectionner un corpus grâce à une combinaison de critères (disciplines, sous-genre textuel, parties textuelles). Le corpus est interrogeable à l'aide de requêtes linguistiques simples ou complexes, à partir des étiquettes morphosyntaxiques ou de relations syntaxiques de dépendance. Des grammaires locales ont été établies sur les thèmes du positionnement et du raisonnement et permettent d'accéder au texte par le biais de recherches sémantiques spécifiques, par exemple sur les verbes d'opinions ou les formulations d'hypothèse.

La problématique linguistique traitée porte sur le positionnement et le raisonnement de l'auteur. A travers le positionnement, l'auteur s'inscrit comme sujet par rapport à ses devanciers, à ses contemporains, il définit sa spécificité, ses choix, comme nous le verrons plus en détail

---

[2] Personnes impliquées au LIDILEM : F. Grossmann, A. Tutin (responsables), G. Antoniadis, F. Boch, C. Cavalla, M. Florez, O. Kraif, I. Novakova, M. Mroué, M.L. Nguyen, F. Rinck.

[3] Personnes impliquées au LLS : J. Osborne, A. Henderson, R. Barr.

[4] Participants : Geoffrey Williams, Chrystel Million.

dans la deuxième section. L'étude du raisonnement permet de retracer son cheminement intellectuel, ce sur quoi il s'appuie et les déductions qu'il opère. Nous souhaitons en outre étudier les différents types de variation que l'on observe en fonction du genre textuel et de la discipline, des disparités importantes ayant été observées pour ce second paramètre. Fløttum *et al*. (2007), en étudiant les articles de recherche en médecine, linguistique et économie dans trois langues (anglais, français et norvégien) ont ainsi mis en évidence que le paramètre disciplinaire était plus déterminant que la langue ou la culture du chercheur. A l'aide de notre corpus et des outils développés, nous souhaitons étendre ces observations à d'autres disciplines, d'autres genres textuels et d'autres paramètres linguistiques.

1.2 Constitution du corpus français d'écrits scientifiques

Le projet Scientext intègre trois grands corpus :

- Un corpus d'écrits scientifiques du français, pluridisciplinaire, et représentant des genres variés, qui contient un peu moins de 5 millions de mots.
- Un corpus anglais d'apprenants, comprenant des travaux longs d'étudiants en anglais langue étrangère (1,1 million de mots).
- Un corpus anglais d'écrits scientifiques, tiré du corpus BMC, principalement en biologie et en médecine, qui avoisine 13 millions de mots, qui a fait l'objet d'études lexicologiques (Williams & Millon, à paraître).

Seul le corpus français sera ici décrit en détail, mais les autres corpus sont annotés selon les mêmes principes (Cf. Henderson *et al.* 2009). Pour étudier les points linguistiques que nous souhaitions explorer, le raisonnement et le positionnement de l'auteur, nous avons constitué pour le français un corpus d'écrits scientifiques diversifié, aussi bien en ce qui concerne le sous-genre (articles scientifiques, communications écrites, thèses ou mémoires d'habilitation à diriger des recherches) que les disciplines. Il était bien entendu exclu d'inclure dans le présent projet la totalité ou la quasi-totalité des disciplines représentées, par exemple par les différentes sections du CNRS[5] ou des sections du CNU[6]. Nous avons ainsi sélectionné des disciplines qui nous paraissaient représentatives de familles scientifiques plus larges et pour lesquelles les écrits étaient facilement disponibles. Trois familles de disciplines sont incluses : les sciences humaines (linguistique, psychologie, sciences de l'éducation et traitement automatique des langues), les sciences expérimentales (biologie, médecine) et les sciences appliquées ou sciences pour l'ingénieur (électronique, mécanique). Les sous-genres sélectionnés intègrent des articles de recherche, des communications écrites[7], des thèses de doctorat et des mémoires d'habilitation à

[5] Liste des sections du Centre National de la Recherche Scientifique : http://www.cnrs.fr/comitenational/sections/intitsec.htm.

[6] Groupes et sections du CNU : http://www.cpcnu.fr/sectionsCnu.htm.

[7] La liste complète des revues et des conférences est donnée en annexe.

diriger les recherches[8]. Le corpus public, consultable en ligne, compte à peu près 5 millions de mots[9]. Le tableau 1 présente le détail du corpus dont on peut relever immédiatement qu'il n'est pas véritablement équilibré : les sciences humaines y sont surreprésentées, en particulier pour les articles, genre absent pour les sciences appliquées où la langue dominante est l'anglais. Les articles obtenus en médecine et biologie sont extraits d'une revue de très bonne qualité *Médecine/Science* qui, sans être une revue de vulgarisation, a néanmoins pour objectif de diffuser au plus grand nombre les recherches récentes dans ce domaine dans la francophonie.

	Articles et communications écrites	Thèses de doctorat	Mémoires d'Habilitation à Diriger des Recherches
Linguistique	67 textes	8 textes	4 textes
Psychologie	12 textes	5 textes	1 texte
Sciences de l'Education	77 textes	6 textes	1 texte
Traitement Automatique des Langues	13 textes	4 textes	
Total Sciences Humaines	169extes	23 textes	6 textes
Biologie	10 textes	11 textes	
Médecine	12 textes	2 textes	
Total Sciences expérimentales	22 textes	13 textes	
Electronique		5 textes	
Mécanique		2 textes	
Total Sciences Expérimentales		7 textes	

Tableau 1 : La composition du corpus français public dans Scientext

Les corpus ont été annotés au plan structurel en suivant les recommandations de la Text Encoding Initiative (TEI Lite, P5[10]), en isolant les différentes parties textuelles de l'article : résumé, introduction, corps du texte, conclusion, remerciements, notes de bas de page, bibliographie, annexes, titres. Ce travail de balisage a été automatisé lorsque cela apparaissait possible, mais la tâche a souvent dû être complétée manuellement, ce qui s'est révélé extrêmement fastidieux, et a nécessité le recrutement de nombreux vacataires qui ont utilisé à

[8] La liste complète des textes ne peut pas être indiquée ici, mais elle est consultable en ligne sur le site de scientext.
[9] Une autre partie du corpus, pour laquelle les droits n'ont pas été obtenus, est consultable sur notre Intranet, avec un mot de passe.
[10] http://www.tei-c.org/Guidelines/P5/.

cette fin des outils spécialisés[11]. L'annotation des parties textuelles est très utile pour mener des études linguistiques fines sur les résumés, les introductions ou les notes de bas de page qui présentent des spécificités manifestes en ce qui concerne le positionnement et le raisonnement. La mise en forme (gras, italique, structure de liste) a été conservée lorsqu'elle pouvait être générée automatiquement, mais pas de façon systématique car le balisage a souvent été réalisé à partir d'un format texte où ces informations avaient été effacées[12]. En outre, le corpus a été analysé linguistiquement (et automatiquement) à l'aide de l'analyseur syntaxique de dépendance Syntex, développé par Didier Bourigault (2007)[13], dont les performances ont été soulignées dans la campagne d'évaluation EASY[14]. Pour chaque phrase, ont été indiqués pour chaque mot le lemme, la catégorie syntaxique et les relations de dépendance qui le lient aux autres mots de la phrase. La figure 1 montre ainsi un exemple pour l'analyse de la phrase : *Enfin, nous avons fait l'hypothèse que les élèves n'occupaient pas une place aléatoire au sein des 4 classes et qu'elle était en relation avec leur statut scolaire.*

[11] En particulier, le logiciel Oxygen : http://www.oxygenxml.com.

[12] Une partie du corpus annotée au plan structurel est disponible pour des fins de recherche. Une convention « creative commons » a été signée à cette fin avec les auteurs et les éditeurs qui ont accepté de rendre leurs corpus disponible. Pour obtenir le corpus, il faut contacter les responsables du projet (scientext@u-grenoble3.fr) et signer une convention.

[13] Que nous remercions très chaleureusement ici.

[14] Voir les détails sur : http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=bourigault&subURL=syntex.html

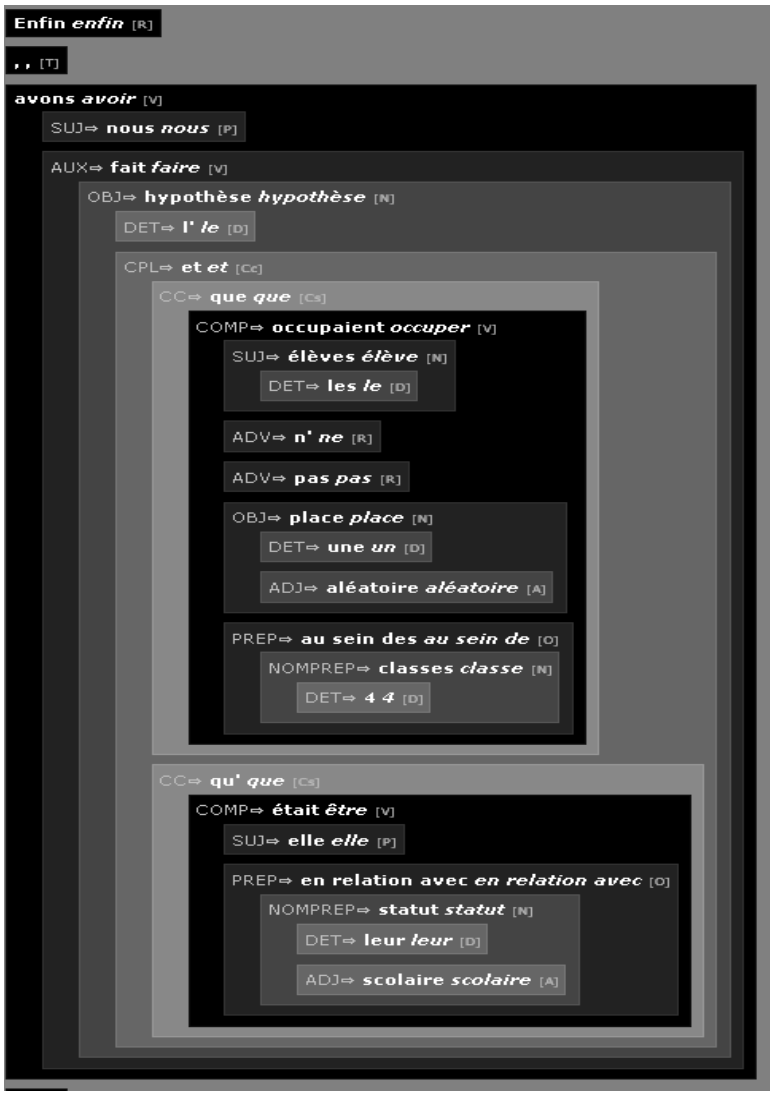


Figure 1 : Analyse syntaxique à l'aide de *Syntex* (Bourigault 2007) de *Enfin, nous avons fait l'hypothèse que les élèves n'occupaient pas une place aléatoire au sein des 4 classes et qu'elle était en relation avec leur statut scolaire*

On repère dans cette analyse à côté des mots sous leur forme fléchie, les lemmes, les catégories syntaxiques et une analyse de dépendance de surface entre les éléments. Par exemple, l'auxiliaire *a* est relié à *fait* par une relation *aux*. Pour une approche plus sémantique (Cf section 3.2.1), il faudra donc recalculer la relation syntaxique « profonde » entre *on* et *fait* dans *on a fait* à partir des relations de surface. L'utilisation de cet analyseur syntaxique permet néanmoins de créer des requêtes et des grammaires générant peu de bruit et peu de silence, en tout cas bien plus performantes qu'un simple étiquetage morpho-syntaxique.

2. Le positionnement dans l'écrit scientifique

2.1 Comment définir le positionnement ?

Un des deux thèmes linguistiques développés dans notre projet est celui du positionnement de l'auteur dans les écrits scientifiques[15]. Le positionnement n'est pas un concept linguistique, contrairement à la notion de « point de vue », utilisée dans les linguistiques de l'énonciation dans le cadre de l'étude de la polyphonie énonciative (par exemple, Nølke *et al.* 2004 ; Rabatel 1988). Dans le cadre de Scientext, nous définissons le positionnement comme la façon dont l'auteur s'inscrit dans une communauté de discours, comment il s'évalue et évalue ses pairs, et quelles propositions propres il met en avant. Ce thème permet d'aborder la figure de l'auteur dans ce type d'écrits, ainsi, plus largement que celle de l'*auctorialité scientifique* à travers l'étude de marques énonciatives, lexicales et syntaxiques spécifiques.

L'étude des marques linguistiques du positionnement dans les écrits scientifiques permet donc d'embrasser trois aspects différents, bien que complémentaires :

a) La question du *contexte scientifique, du cadre théorique et des références* propres à un auteur ou à une équipe. Il peut s'agir de la filiation intellectuelle, c'est-à-dire l'approche, les idées, voire la terminologie dont un auteur s'inspire, ou le cadre théorique dans lequel il s'inscrit explicitement (Boch. F *et al*. 2007; Garcia P.P. 2008; Grossmann *et al.* 2009). Cette problématique inclut également l'ensemble des références à autrui « neutres » qui apparaissent dans l'écrit scientifique, même si le positionnement de l'auteur par rapport aux auteurs mentionnés apparaît ici moins explicite[16].
b) Dans le sens plus restreint de « prise de position », l'étude du positionnement permet d'observer *les moyens linguistiques utilisés pour exprimer un parti pris, un jugement ou une évaluation*. Cela peut concerner l'évaluation d'un point théorique, d'un résultat, d'une démarche (Cf. Tutin 2010a, Cavalla & Tutin, à paraître), sur la démarcation ou la distance vis à vis des pairs (*contrairement à X … nous nous différencions de X*) (Chavez 2008), ou bien au contraire l'adhésion ou la convergence de vue. L'évaluation peut aussi concerner la

[15] La question du raisonnement ne sera pas abordée dans cet article.

[16] Dans le cadre du projet Scientext, nous distinguons la citation ou la référence à autrui « neutre » qui n'indique pas de positionnement explicite de l'auteur par rapport à la référence citée, de la citation ou référence à autrui « positionnée » qui indique une position explicite de l'auteur (Ex : *contrairement à Duschmoll (1970) … Nous reprenons le modèle de Duschmoll (1970)*) … La citation neutre n'est généralement pas intégrée syntaxiquement comme en (1), alors que la citation positionnée est intégrée syntaxiquement comme en (2) (Cf. Florez, 2010) :

(1) Les travaux sur le positionnement dans les écrits scientifiques sont souvent d'inspiration énonciative (Cf. Fløttum *et al.* 2006).

(2) Contrairement à Fløttum (2006), notre approche du positionnement n'est pas exclusivement énonciative.

Bien entendu, toutes les citations intégrées ne sont pas positionnées.

démarche de l'auteur, dans la formulation de la conformité/non-conformité aux attentes (Rinck *et al.* 2007) ;

c) Enfin, le positionnement concerne les *choix propres, les propositions et les déductions opérés* par l'auteur, par le biais de l'emploi de la première personne (*nous optons pour … nous concluons que … nous avons montré que…*) ou l'emploi d'un lexique déontique (*il faut … ce problème doit à nouveau être traité …*).

Dans ce projet, nous avons fait l'hypothèse que l'expression du positionnement était relativement stéréotypée et s'exprimait à travers une phraséologie repérable, qui pouvait être traitée dans des grammaires locales (de l'opinion, de l'évaluation, de la démarcation …). Nous avons aussi supposé que ces expressions récurrentes appartenaient à un sous-langage plus spécifique du genre que des disciplines en tant que telles, un lexique scientifique transdisciplinaire (Tutin 2007) de la langue scientifique générale au sens de Pecman (2004).

2.2 L'exemple du lexique verbal du positionnement

Plusieurs études ont été réalisées sur le thème du positionnement dans le cadre du projet Scientext. Nous reprenons ici un exemple représentatif des études sur ce thème, l'utilisation du lexique verbal marquant explicitement un engagement de l'auteur (Tutin, 2010b). Les écrits scientifiques sont souvent considérés comme un genre « neutre », avec un fort effacement énonciatif, où l'auteur se dissimulerait derrière la présentation de faits objectifs et des modalités de raisonnement partagés par la communauté scientifique. Les travaux accomplis sur ce sujet dans les dernières années (par exemple, Swales 1990 ; Hyland 2002 ; Fløttum *et al*. 2006 ; Rinck 2006) montrent cependant qu'il n'en est rien, en tout cas dans certaines disciplines, et que l'écrit scientifique est véritablement un texte argumentatif où la dimension rhétorique est fortement présente. Cette étude, qui comparait trois disciplines des sciences humaines et sociales, la linguistique, la psychologie et les sciences de l'éducation[17], cherchait à mettre en évidence les modalités d'engagement explicites de l'auteur à travers les verbes de positionnement associés à un pronom sujet (Ex : *je cherche à démontrer, nous pensons que..*). Nous faisions l'hypothèse que la présence auctoriale s'établit diversement selon les disciplines des sciences humaines. On peut imaginer, comme mis en évidence par Fløttum *et* al. (2006), qu'elle sera assez manifeste en sciences du langage où l'auteur cherche souvent à développer une pensée ou un modèle propre. En psychologie cognitive et sociale, en revanche, on peut supposer que les écrits, qui se rapprochent par la structure IMRaD[18] et les méthodes (expérimentales) des sciences dures, pourraient ainsi en adopter le style plus « objectif », avec moins de références explicites aux auteurs de l'article et l'emploi moins marqué de verbes exprimant un point de vue explicite. En

[17] D'autres disciplines sont abordées dans le cadre du projet Scientext, mais nous avons choisi de nous limiter ici à un sous-ensemble de sciences humaines et sociales, Fløttum *et al*. (2006) ayant déjà montré que la présence auctoriale était faible dans les sciences expérimentales comme la médecine.

[18] IMRaD : Introduction, Méthodes, Résultats, Analyse, Discussion. C'est un plan textuel imposé dans les disciplines expérimentales.

outre, on peut aussi s'attendre à ce que le type de verbe utilisé soit fortement lié à la valeur référentielle du pronom sujet, selon qu'il renvoie strictement à l'auteur ou qu'il intègre aussi la communauté de discours.

L'étude a été réalisée partir d'un corpus de 60 articles de linguistique, psychologie et sciences de l'éducation[19] (3x20), en observant de façon systématique dans les introductions et les conclusions les verbes qui engagent fortement l'auteur, associés à un pronom auteur repéré semi-automatiquement (*je, nous, on*).

Ont été retenus comme verbes de positionnement explicites :

- des verbes qui expriment une **opinion** ou un **point de vue** (*penser, croire, considérer que, juger…*), ou une distance/adhésion par rapport aux pairs (*se distinguer de, rejoindre…*), ou à un questionnement (*se demander…*)
- des verbes indiquant un **choix** (*choisir, retenir, opter pour…*), une **intention** (*vouloir, souhaiter, projeter…*) ou des hypothèses (*faire, formuler, émettre une hypothèse ; supposer*).
- des verbes qui indiquent un apport **apport spécifique de l'auteur**, une **proposition** (*proposer …*), une **preuve** ou une **démonstration** (*montrer, prouver …*) ou bien des **résultats** (*dégager, souligner …*).

Les résultats de cette étude ont permis de mettre en évidence deux grandes tendances. La première est avant tout une visibilité assez modérée de l'auteur. S'il ne se cache pas (il apparaît à la première personne, voire même à la première personne du singulier en linguistique), dans les textes examinés, l'auteur se manifeste discrètement. Ainsi, le *nous* de modestie est partout préféré. Bien que ce *nous* conventionnel soit souvent difficile à interpréter de façon univoque, l'auteur semble mettre par cet emploi l'accent sur son appartenance à une communauté de discours : il met peu en avant son individualité et sa spécificité. En outre, les verbes de positionnement employés ne sont pas majoritairement des verbes à « fort » positionnement (comme les verbes d'opinion), comme on l'observe sur la figure 2, mais plutôt des verbes qui indiquent les choix effectués par l'auteur (Ex : *nous avons opté …*) ou les apports scientifiques de la recherche effectuée (Ex : *Nous avons dégagé …*). Enfin, on relève que les verbes à « fort positionnement, comme *penser*, tendent à être modalisés dans des formules du type *on peut penser que …* On pourrait voir dans ce type de modalisation une prise de risque minimale de l'auteur, qui ne cherche pas ici à s'engager dans une opinion affirmée, mais on peut surtout interpréter ce type de formulation, à l'instar de Hyland (1998), comme une forme de négociation avec le lecteur (le *on* inclut ici le lecteur). L'auteur ne cherche pas ici à imposer son point de vue (ce n'est pas le mode de fonctionnement de la « Science ») mais il montre qu'au vu des résultats obtenus, tout chercheur (y compris l'auteur et le lecteur), tirerait des conclusions identiques. En bref, ces éléments

[19]Certains des articles analysés ici n'appartiennent pas au corpus public présenté en 1.

semblent montrer que l'auteur s'inscrit avant tout fortement dans la communauté de discours, et met peu en avant son individualité.

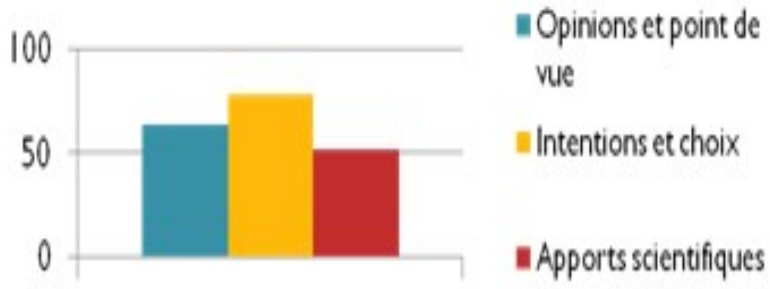


Fig. 2 : Répartition des différents types de verbes de positionnement

La deuxième grande tendance observée est, comme le notent Fløttum *et al.* (2006), une forte variabilité disciplinaire au sein de ces trois disciplines des sciences humaines, en tout cas dans ce corpus.
Comme on l'observe dans la figure 3, la proportion des verbes de positionnement va ainsi de 1 pour la psychologie à 3 pour la linguistique.

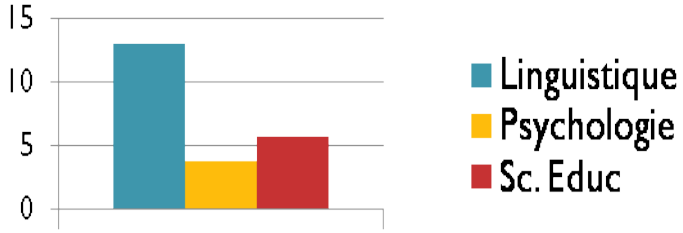


Fig. 3 : Répartition disciplinaire des verbes de positionnement

Si l'on entre davantage dans les détails, on s'aperçoit que la linguistique se caractérise par une forte proportion de verbes de positionnement, de toutes sortes, mais en particulier d'opinion, intentions, résultats et démonstration, ainsi que par la présence notable du *je* (que l'on rencontre peu ailleurs). En sciences de l'éducation, la présence des verbes de positionnement est plus modérée, et l'accent est mis sur la justification de la démarche, à travers les opinions et les intentions. En psychologie enfin, on dénombre encore moins de verbes de positionnement et ceux-ci apparaissent surtout pour indiquer une opinion et une démarche expérimentale.
On voit ici que l'étude linguistique menée, qui doit bien sûr être complétée, permet de mettre en évidence certains éléments essentiels du fonctionnement des disciplines, comme les critères de

scientificité et d'évaluation propres à chacune. Ainsi, la linguistique semble mettre l'accent sur une forme d'individualité et de créativité de l'auteur, alors qu'en sciences de l'éducation, les raisons d'être (sociales ?) de la recherche sont soulignées. La psychologie, qui se rapproche sur ce plan des sciences expérimentales, met plutôt l'accent sur les hypothèses et les résultats obtenus. Il faut cependant se garder de conclusions trop hâtives : les études linguistiques ont montré que la question du positionnement était plurielle et que toutes ses dimensions ne convergeaient pas nécessairement. Ainsi, si la linguistique se caractérise par une certaine visibilité de l'auteur, les écrits de cette discipline présentent par ailleurs peu de marqueurs de positionnement par rapport aux pairs. En économie, en revanche, on observe la tendance inverse avec des stratégies de persuasion plus offensives en direction de pairs à travers les marques d'évaluation (Tutin 2010) ou de filiation (Grossmann *et al.* 2009) mais une visibilité de l'auteur moindre. Des études de cas sur le corpus Scientext sur plusieurs points linguistiques (citation positionnée, propositions propres de l'auteur ...) restent à accomplir pour brosser un portait nuancé et adéquat du positionnement.

3. Présentation du site Scientext : modes d'exploitation des corpus

Dans le cadre de notre projet, outre la constitution de corpus, un outil d'exploitation des corpus annotés a été élaboré par Achille Falaise sous la forme d'un site Internet (adresse : http://scientext.msh-alpes.fr), librement consultable, pour interroger les corpus, entre autres, sur les marques du positionnement et du raisonnement à l'aide de grammaires prédéfinies. L'exploitation du corpus se fait en trois étapes. L'utilisateur sélectionne le corpus dans un premier temps, puis effectue sa requête ("sémantique", "libre" et "guidée", ou "avancée"). Enfin, il affiche les résultats sous formes de concordances ou de traitements statistiques simples.

3.1 Sélection des textes

La première étape consiste à sélectionner le corpus, à la façon de Frantext, selon un ensemble de paramètres, comme indiqué sur la copie d'écran figure 4. L'usager peut choisir la ou les discipline(s), les genres textuels et les parties textuelles, grâce au balisage structurel réalisé. Il pourra par exemple sélectionner les résumés des articles et communications des sciences humaines.

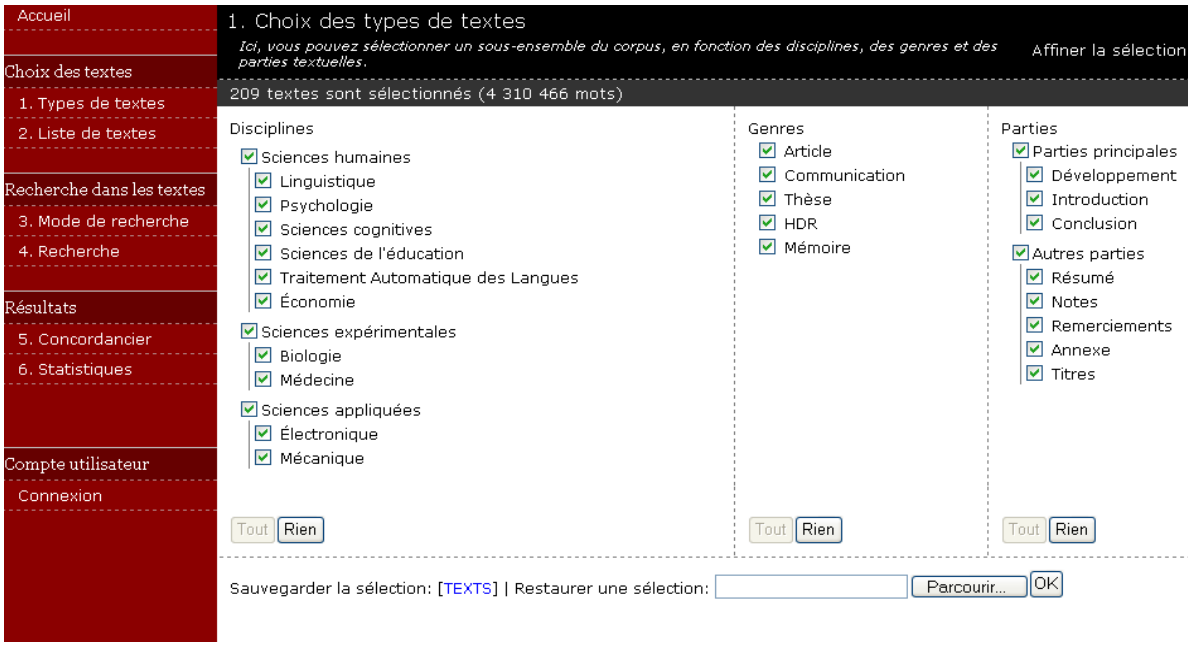


Fig. 4 : Sélection du corpus dans Scientext

Une fois les textes correspondant aux choix affichés, l'usager peut ensuite affiner la sélection en sélectionnant ou non les textes un par un. Il est également possible de mémoriser le corpus sélectionné de façon à le réutiliser dans une session ultérieure.

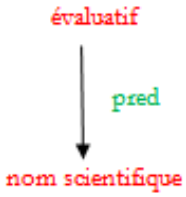
3.2 Recherche dans les textes

Une fois le corpus sélectionné selon les disciplines, les genres textuels et les parties textuelles désirés, l'utilisateur peut accéder au contenu du texte par trois types de recherche : un mode sémantique et guidé, axé sur la question linguistique du positionnement et du raisonnement, un mode de recherche libre et guidé, et un mode avancé utilisant des expressions régulières. Tous ces modes de recherche sont traduits dans le même langage de requête ConcQuest développé par Olivier Kraif (2008) et étendu par Achille Falaise (Falaise & Tutin 2010) dans le cadre du projet Scientext.

3.2.1 Le mode sémantique guidé : accès par les grammaires locales

Le mode sémantique permet de faire des recherches dans les textes à partir de thèmes liés au positionnement ou au raisonnement, comme les opinions ou l'évaluation des objets scientifiques. Ce mode sémantique est construit à partir de schémas sémantico-rhétoriques, qui sont ensuite traduits dans le langage de requête, à l'aide de variables et de dépendances syntaxiques (Cf. Tutin 2010c ; Falaise & Tutin 2010).

A titre d'exemple, voici le schéma simple utilisé pour l'évaluation adjectivale des objets scientifiques, qui associe un élément évaluatif à un nom scientifique.



On fera correspondre à ce schéma une grammaire utilisant pour chaque notion un ensemble de lexèmes (par exemple, pour les noms scientifiques : *méthode, problème, question,* etc. Pour les adjectifs évaluatifs, *adapté, pertinent, original, novateur* ...), alors que la relation pred (prédicat) sera traduite par la relation syntaxique épithète ou attribut. La syntaxe des grammaires, élaborée par Achille Falaise et Olivier Kraif, permet de redéfinir des relations, d'utiliser des variables et des raccourcis d'écriture. La principale difficulté est cependant de traduire des relations syntaxiques de surface en relations sémantiques, ce qui exige une excellente connaissance de l'analyse syntaxique de surface réalisée par le système Syntex (Bourigault 2007).

```
//TITRE: Adjectifs d'évaluation

//INFO: Les adjectifs d'évaluation qui portent sur les noms scientifiques

(ATTRIB,#2,#1) = (SUJ,#3,#1) (ATTS,#3,#2) ;

$eval=acceptable,adéquat,aisé,ambitieux,approximatif,bon,clair,classique,cohérent,complexe,concis,confus,convaincant,correct,crucial,déterminant,difficile,discutable,encourageant,épineux,essentiel,excellent,faible,fin,flou,fondamental,important,innovant,intéressant,irréprochable,judicieux,majeur,mauvais,meilleur,important,pertinent,nouveau,original,passable,passionnant,performant,principal,prometteur,riche,rigoureux,satisfaisant,séduisant,sérieux,significatif,solide,souhaitable,stimulant,vague,valable

$theo=
analyse,approche,article,caractéristique,choix,communication,concept,contribution,critère,élément,étude,exemple,facteur,fonction,idée,méthode,modèle,notion,objectif,phénomène,problème,projet,proposition,qualité,question,réflexion,résultat,solution,test,théorie,travail

Main = <lemma=$eval,#1> && <lemma=$theo,#2> :: (ATTRIB,#1,#2) OR (ADJ,#1,#2);
```

Ces grammaires sont ensuite intégrées dans l'interface et peuvent être librement choisies par l'utilisateur, comme cela apparaît sur la figure 5.

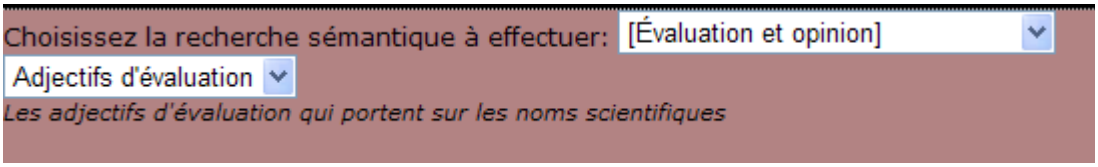


Fig. 5. : Choix des requêtes sémantiques

Un ensemble de grammaires a été élaboré à cette fin autour de différents thèmes : évaluation, opinion, démarcation, auteurs cités, formulation des hypothèses, etc.

3.2.1 Le mode simple et guidé

Ce mode de recherche guidé permet à l'utilisateur de sélectionner des formes, lemmes et/ou catégories, ainsi que les relations syntaxiques désirées. La requête présentée à la figure 6 permet ainsi d'extraire les occurrences où *hypothèse(s)* est le complément d'objet direct d'un verbe, indépendamment de la position des éléments dans la phrase. Cette requête extraira des exemples comme *faire des hypothèses, formuler des hypothèses, confirmer cette hypothèse* ... (Cf. Figure 8).

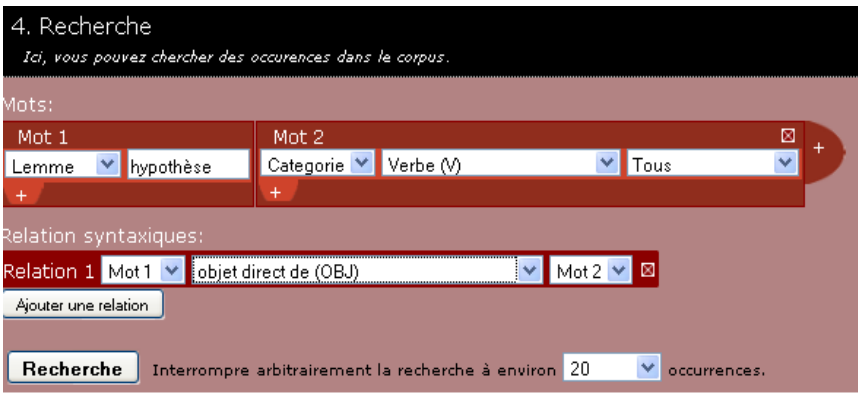


Fig. 6 : un exemple de requête libre et guidée (le nom *hypothèse* objet direct d'un verbe)

3.2.2 Le mode avancé

L'utilisateur à l'aise avec les expressions régulières et la syntaxe de dépendance pourra aussi développer ses propres grammaires. Les grammaires exploitent la linéarité et/ou les dépendances syntaxiques. Elles permettent de redéfinir des relations syntaxiques, afin d'en proposer un traitement plus sémantique. Par exemple, dans la Figure 7, on définit une nouvelle relation SUJCOMP (sujet dans le cas des verbes composés) qui peut apparaître entre le sujet et le participe passé, par exemple dans *nous avons proposé*[20]. Les grammaires intègrent également des

[20] Syntex est un analyseur syntaxique de surface. Dans *nous avons proposé*, il y a ainsi une relation SUJ entre l'auxiliaire et le sujet, et une relation AUX entre l'auxiliaire et le participe passé. A l'aide des redéfinitions de relations, on pourra ainsi proposer des analyses plus sémantiques du corpus de façon à analyser la relation profonde entre *nous* et *proposer* comme dans les exemples suivants : *nous avons pu proposer, nous venons de proposer, nous avons été contraints de proposer ….*

variables pour réaliser des raccourcis d'écriture.

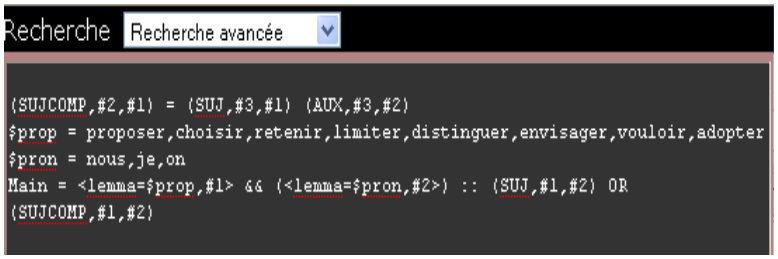


Fig. 7 : Un exemple de recherche en mode avancé

3.3 Affichage et statistiques

Une fois la recherche effectuée, l'utilisateur peut ensuite faire afficher les résultats de plusieurs façons, les exporter afin de les traiter localement ou obtenir des traitements statistiques simples.

3.3.1 Affichage des résultats et exportation

Une fois le corpus sélectionné, on peut l'afficher dans un concordancier KWIC, dont les fenêtres sont paramétrables. La référence du texte, ainsi que la partie textuelle, sont indiquées. L'utilisateur peut également demander un contexte plus large (dans la limite de 200 mots, comme dans la figure 8.

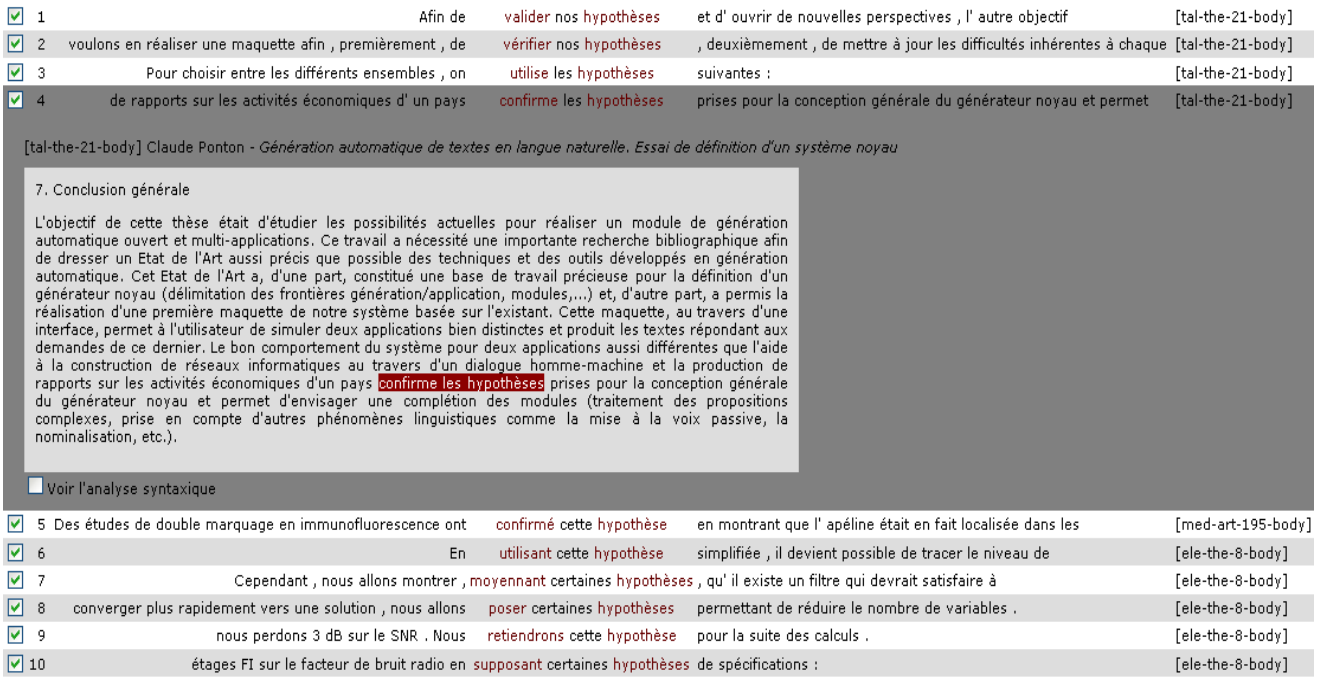


Fig. 8 : Affichage des concordances : concordance KWIC et concordance large

On peut aussi obtenir l'affichage des structures syntaxiques, comme cela apparaît dans la figure 1. L'utilisateur peut également sauvegarder ses résultats, après avoir décoché les résultats inappropriés, dans un fichier html ou un fichier Excel, sur lequel il pourra retravailler ultérieurement.

3.3.2 Statistiques

Des statistiques simples sur les résultats sont également intégrées à l'interface. On peut ainsi obtenir la liste des lemmes correspondant à une requête. La recherche sur les verbes ayant le lemme *hypothèse* comme objet direct (967 occurrences pour la totalité du corpus) montre ainsi que la collocation *faire* DET *hypothèse* est de loin l'expression la plus courante (Cf. Figure 9).

Au total 967 occurrences ont été trouvées.

Quelles statistiques souhaitez-vous consulter ?

Liste des lemmes

Lemme	Nombre	Forme
/faire/ /hypothèse/	242 (25.03%)	faire hypothèse (80) , faisons hypothèse (67) , fait hypothèse (45) , faisant hypothèse (12) , faire hypothèses (9) , font hypothèse (8) , ferons hypothèse (6) , faisions hypothèse (4) , faisons hypothèses (2) , Faites hypothèses (2) , fait hypothèses (1) , fasse hypothèse (1) , fais hypothèses (1) , fais hypothèse (1) , fit hypothèse (1) , faites hypothèses (1) , faisais hypothèse (1)
/tester/ /hypothèse/	80 (8.27%)	tester hypothèse (42) , tester hypothèses (17) , testé hypothèse (5) , testant hypothèses (3) , testerons hypothèse (3) , testons hypothèse (2) , testant hypothèse (2) , teste hypothèses (2) , teste hypothèse (1) , testent hypothèses (1) , testé hypothèses (1) , Tester hypothèse (1)
/confirmer/ /hypothèse/	64 (6.62%)	confirmer hypothèse (21) , confirment hypothèse (18) , confirme hypothèse (10) , confirmer hypothèses (4) , confirmé hypothèse (4) , confirment hypothèses (3) , confirmant hypothèse (2) , confirme hypothèses (1) , confirmerait hypothèse (1)
/formuler/ /hypothèse/	55 (5.69%)	formuler hypothèse (21) , formuler hypothèses (13) , formulons hypothèse (11) , formulé hypothèse (6) , formule hypothèse (2) , formulent hypothèses (1) , formulent hypothèse (1)
/émettre/ /hypothèse/	55 (5.69%)	émettre hypothèse (22) , émis hypothèse (8) , émettre hypothèses (6) , émettons hypothèse (5) , émet hypothèse (4) , émettent hypothèse (4) , émettons hypothèses (2) , émettrons hypothèse (1) , émis hypothèses (1) , émettent hypothèses (1) , émettant hypothèse (1)
/vérifier/ /hypothèse/	52 (5.38%)	vérifier hypothèse (37) , vérifier hypothèses (4) , vérifié hypothèse (3) , vérifient hypothèse (2) , vérifie hypothèse (2) , vérifient hypothèses (1) , vérifie hypothèses (1) , vérifions hypothèse (1) , vérifierons hypothèses (1)
/valider/ /hypothèse/	33 (3.41%)	valider hypothèse (14) , valider hypothèses (7) , valide hypothèse (4) , valident hypothèse (3) , validant hypothèse (2) , valident hypothèses (1) , validant hypothèses (1) , valide hypothèses (1)
/poser/ /hypothèse/	33 (3.41%)	poser hypothèse (8) , posons hypothèse (7) , poser hypothèses (7) , poserons hypothèses (2) , posé hypothèse (2) , poserons hypothèse (1) , posions hypothèse (1) , pose hypothèses (1) , posons hypothèses (1) , pose hypothèse (1) , posant hypothèse (1) , posé hypothèses (1)
/proposer/ /hypothèse/	28 (2.90%)	proposer hypothèses (9) , proposer hypothèse (7) , proposerons hypothèses (3) , proposé hypothèse (2) , propose hypothèse (2) , proposé hypothèses (1) , propose hypothèses (1) , proposent hypothèse (1) , proposant hypothèses (1) , proposons hypothèses (1)

Fig. 9 : Les structures V-OBJ-> *hypothèse* les plus fréquentes

Il est également possible d'obtenir la répartition de la réponse dans les disciplines, les genres textuels ou les parties textuelles, en obtenant les fréquences absolues ou les fréquences relatives[21]. Ainsi, la grammaire des verbes d'opinion appliquée à la totalité du corpus (Cf. Figure 10) montre que ce type de verbe est fort fréquent dans les remerciements et les conclusions, mais moins usuel dans les introductions, notes et résumés (ce sont les fréquences relatives qui sont ici prises en considération).

[21] Fréquence relative : calcul du nombre d'occurrences sur le nombre total d'occurrences du texte.

Développement	1663	/	3645711	=	4.56 ‰
Introduction	69	/	209266	=	3.3 ‰
Conclusion	54	/	85200	=	6.34 ‰
Notes	49	/	153355	=	3.2 ‰
Annexe	24	/	118198	=	2.03 ‰
Remerciements	14	/	17551	=	7.98 ‰
Résumé	7	/	25061	=	2.79 ‰

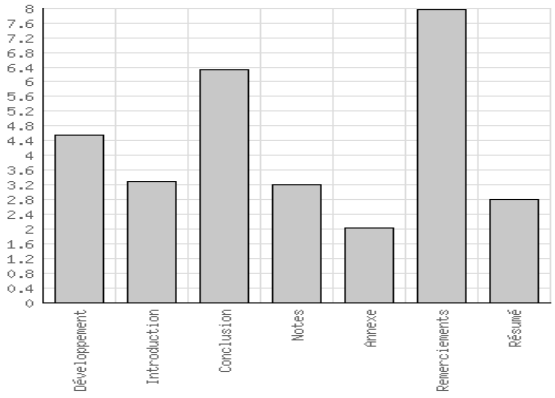


Fig. 10 : Répartition des verbes d'opinion dans le corpus Scientext selon la partie textuelle

En ce qui concerne le genre textuel (Cf. Fig. 11), on observe d'intéressantes différences dans notre corpus. Si, de façon attendue, c'est dans les mémoires d'HDR que l'on trouve le plus d'expressions de l'opinion (verbale), on relève une intéressante différence entre les articles et les communications écrites qui sont souvent assimilés. La communication écrite serait-elle plus proche du genre oral, avec un engagement de l'auteur plus explicite ?

Au total 1557 occurrences ont été trouvées.

Quelles statistiques souhaitez-vous consulter ?

Répartition des lemmes

Genre	Nombre absolu d'occurrences		Nombre de mots total		Nombre relatif d'occurrences
Thèse	1146	/	3665882	=	3.13 ‰
HDR	188	/	454961	=	4.13 ‰
Communication	185	/	487188	=	3.8 ‰
Article	38	/	231903	=	1.64 ‰

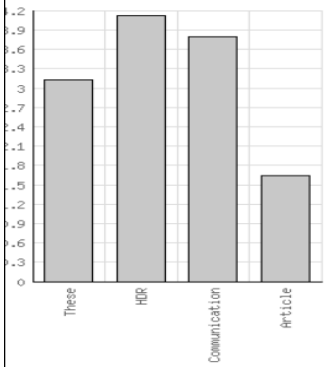


Fig. 11 : Répartition des verbes d'opinion par genre textuel

Enfin, il est également possible d'obtenir la répartition des réponses par discipline.

Ces statistiques simples permettent de fournir rapidement des éléments sur la comparaison textuelle et disciplinaire. Ces résultats chiffrés ne doivent cependant pas être interprétés trop hâtivement et un retour au contexte sera souvent nécessaire pour désambiguïser une interprétation et analyser plus en finesse les formulations linguistiques.

Conclusion

Le projet Scientext est à la fois un projet ingénierique, avec l'élaboration d'un corpus d'écrits scientifiques diversifié librement mis à la disposition de la communauté linguistique à l'aide d'outils logiciels, et un projet de linguistique théorique visant à mieux comprendre le fonctionnement linguistique du positionnement et du raisonnement de l'auteur scientifique. Le premier volet est désormais réalisé, avec la mise à disposition d'un site web opérationnel, alors que le second volet, qui a fait l'objet d'un ensemble d'études sur la filiation, l'évaluation, la démarcation, les verbes de positionnement, le raisonnement causal, doit encore être développé. Ces premières études linguistiques permettent de brosser un portrait nuancé et diversifié de la question du positionnement de l'auteur, qui revêt plusieurs facettes qui ne sont pas nécessairement convergentes.

Nous espérons que le corpus Scientext qui, à notre connaissance, est un des seuls corpus analysés syntaxiquement librement consultables en ligne[22], sera largement exploité par la communauté des linguistes et sera suivi d'autres projets de ressources textuelles libres permettant de développer la linguistique de corpus en France.

Bibliographie

Boch, Françoise, Francis Grossmann, Fanny Rinck (à paraître). Le cadrage théorique dans l'article scientifique: Un lieu propice à la circulation des discours, *Actes du colloque international Cit-dit*, Circulation des discours et liens sociaux: Le discours rapporté comme pratique sociale, Québec, du 5 au 7 octobre 2006, Nota Bene.

Bourigault, D. (2007). *Un analyseur syntaxique opérationnel : SYNTEX*. Thèse d'habilitation à diriger des recherches. Université Toulouse le Mirail.

Cavalla, C ., Tutin, A. (à paraitre). Etude des collocations évaluatives dans les écrits scientifiques. In L. Gautier et S. Méjri (Eds). *Les collocations dans les discours spécialisés.* Dijon, Editions Universitaires de Dijon.

Chavez Ingrid (2008), *La démarcation dans les écrits scientifiques - Les collocations transdisciplinaires comme aide à l'écrit universitaire auprès des étudiants étrangers*, Mémoire de Master Français Langue Etrangère Recherche, ss.dir. Cristelle Cavalla, Université Stendhal-Grenoble3. Grenoble.

Falaise, A., Tutin, A. (2010). Approche onomasiologique de la phraséologie transdisciplinaire des écrits scientifiques : la recherche sémantique dans les textes dans le cadre du projet Scientext, Démonstration, *Conférence THot. 4-5 juin 2010, Annecy*.

Florez, M. (2010). Marques de la citation positionnée dans trois disciplines des sciences humaines. *Colloque international des Etudiants chercheurs en Didactique des Langues et en Linguistique* Du 29 juin au 2 juillet

[22] Avec Corpus Eye, analysé avec le système VISL : http://beta.visl.sdu.dk/

2010, Université Stendhal, Grenoble, France.

Fløttum, K., Dahl, T. & Kinn, T. (2006). *Academic Voices across languages and disciplines*. Amsterdam/Philadelphia, John Benjamins.

Garcia da Silva, P. P. (2008). *Les marques de la filiation dans les écrits scientifiques*. Mémoire de Master 1, sous la direction de Francis Grossmann et d'Agnès Tutin, Université Stendhal-Grenoble.

Grossmann, F.., Tutin A., Garcia da Silva P. (2009). Filiation et transferts d'objets scientifiques dans les écrits de recherche, *Pratiques* 143-144, 187-202.

Henderson, A., Tutin, A., Grossmann, F., Barr, R. (2009). SCIENTEXT : A Corpus of French and English Scientific Texts. *British Association of Applied Linguistics Annual Conference*, 4 septembre 2009, Newcastle University.

Hyland K. (2002). Authority and invisibility : authorial identity in academic writing. *Journal of Pragmatics*. Vol 34, 8. 1091-1112.

Hyland, K. (1998). *Hedging in scientific research articles*. Amsterdam/Philadelphia, John Benjamins.

Kraif, O. & Tutin, A. (2009). Using a bilingual annotated corpus as a writing aid: An application for academic writing for EFL users. In N. Kübler (Ed.) Proceedings of TaLC7, 7ème Conference Teaching and Language Corpora. Bruxelles: Peter Lang.

Kraif, O. (2008). Comment allier la puissance du TAL et la simplicité d'utilisation ? l'exemple du concordancier bilingue ConcQuest, *Actes des 9ème Journées d'analyse statistique des données textuelles, JADT 2008*. Lyon, Presses universitaires de Lyo.: 625-634.

Nølke, H., Fløttum, K. & Norén, C. (2004). *ScaPoLine. La théorie scandinave de la polyphonie linguistique.* Paris : Editions Kimé.

Rabatel, A. (1998). *La construction textuelle du point de vue*. Lausanne/Paris : Delachaux et Niestlé.

Rinck, F. (2006). *L'article de recherche en Sciences du Langage et en Lettres, Figure de l'auteur et approche disciplinaire du genre*. Thèse de doctorat en Sciences du Langage, sous la direction de F. Boch et F. Grossmann, Université de Grenoble.

Siddharthan A., S. Teufel. (2007). Whose idea was this, and why does it matter? Attributing scientific work to citations. In: "Proceedings of NAACL/HLT-07", Rochester, New York.

Tutin, A. (2007) (coord.), Lexique et écrits scientifique, *Revue Française de Linguistique Appliquée*, volume XII-2, décembre 2007.

Tutin, A. (2010a). Evaluative adjectives in academic writing in the humanities and social sciences. *Interpersonality in written academic discourse: perspectives across languages and cultures*. Cambridge : Cambridge Publishing. 219-239.

Tutin, A. (2010b). *Dans cet article, nous souhaitons montrer que ...* Lexique verbal et positionnement de l'auteur dans les articles en sciences humaines. Enonciation et rhétorique dans l'écrit scientifique. *LIDIL 41.*

Tutin, A. (2010c). Showing phraseology in context: an onomasiological access to lexico-grammatical patterns in corpora of French scientific writings, *Proceedings of eLexicography in the 21st century: new challenges, new applications, 22-24 october 2009, Louvain la Neuve.*

Williams, G., Millon, Chr. (à paraître 2010). The General and the Specific : Collocational resonance of scientific language. *Proceedings Corpus Linguistics 2009*. University of Liverpool