INTRODUCTION

Francis Grossmann, Agnès Tutin

L'ouvrage que nous présentons ici veut rendre compte de travaux effectués à partir d'un projet de recherche spécifique, le projet Scientext, développé par une équipe de linguistes dans le but de mieux connaître certaines des caractéristiques des discours scientifiques écrits contemporains, dans différentes disciplines et à travers deux langues principales, le français et l'anglais. Trois équipes ont été impliquées dans ce projet. Le LIDILEM[1] (Laboratoire de Linguistique et de Didactique du Française Langue Étrangère 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[2], 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[3]), a recueilli et traité un corpus d'anglais scientifique, principalement en sciences de la vie et médecine.

Dans cette introduction, nous présentons dans un premier temps une synthèse des études abordées dans le cadre de Scientext et leurs spécificités par rapport aux autres travaux sur les écrits scientifiques. Les outils méthodologiques développés dans le cadre du projet sont ensuite abordés dans la deuxième section. Nous finissons par une brève présentation des contributions de ce volume.

1. Scientext et les travaux sur les écrits scientifiques

1. 1. Position du projet dans les études sur la science

Si l'étude de la langue et de la rhétorique scientifique a déjà fait l'objet de nombreuses investigations (Rinck, 2010, pour une synthèse récente), la jonction entre le versant proprement linguistique et discursif et les versants épistémologique et sociologique, au cœur des courants d'étude consacrés aux conditions de l'activité scientifique, reste assez

1. 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.
2. Personnes impliquées au LLS : J. Osborne, A. Henderson, R. Barr.
3. Participants : Geoffrey Williams, Chrystel Millon.

« L'écrit scientifique : du lexique au discours », Agnès Tutin et Francis Grossmann (dir.)
ISBN 978-2-7535-2846-8 Presses universitaires de Rennes, 2014, www.pur-editions.fr

peu explorée. Dans le champ linguistique, et notamment dans le cadre des langues de spécialité ou de l'« écrit académique » (voir par ex. les travaux menés dans le cadre de l'ESP – l'English for specific purposes, ou de l'EAP – English for Academic Purposes) ou, plus récemment, l'aide à la rédaction en anglais (Pecman, 2004), les recherches sont souvent restées coupées de celles effectuées dans le champ plus général des études de la science (« Sciences Studies »), tant dans leur versant épistémologique que sociologique[4]. De même, l'intérêt porté aux dispositifs textuels et matériels de l'écrit scientifique est resté relativement faible dans l'espace francophone, et les travaux les plus intéressants en la matière, si l'on excepte les travaux portant sur les dispositifs sémiotiques de la vulgarisation scientifique (Beacco et Moirand, 1995 ; Jacobi et Schiele, 1988 ; Jacobi, 1999 ; Jeanneret, 1994), sont le fait d'historiens du livre et de l'éditorialisation (Chartier, 1992), de sociologues (Pontille, 2001) collaborant avec des anthropologues de l'écrit/linguistes (Franckael, 1992).

En France, les analystes de discours à la suite de Michel Pêcheux, se sont intéressés d'abord aux discours politiques, vecteurs d'idéologies, ou fondant la mémoire collective (Tournier, 1986 ; Guilhaumou et Maldidier, 1986). C'est plutôt dans le secteur de ce que l'on appelait alors la « linguistique appliquée », que certaines formes d'écrit scientifique ont suscité par la suite des analyses, en particulier les discours de vulgarisation, proches des préoccupations didactiques (Peytard *et al.*, 1984 ; Jacobi, 1986 ; Loffler-Laurian, 1993 ; Beacco et Moirand, 1995). L'intérêt pour les discours universitaires, et plus précisément pour les écrits de recherche (Dabène et Reuter, 1998) s'est progressivement accru, d'autant que la massification de l'enseignement supérieur exigeait une attention nouvelle aux difficultés rencontrées par les étudiants, confrontés aux spécificités des genres scientifiques, tant en réception qu'en production. En linguistique française, certains travaux se sont également intéressés aux spécificités de l'écrit scientifique. Dès 1992, Dominique Maingueneau avait signalé le « tour ethnolinguistique » pris par l'analyse de discours, et opéré de fructueuses jonctions entre des notions clés issues de Michel Foucault (*autorité, auctorialité, légitimité*) et celles de la linguistique de l'énonciation. Il est à noter que chez les linguistes et analystes du discours, des objets jusque-là ignorés ont également fait l'objet de recherches, à partir des outils fournis par la pragmatique et la linguistique interactionnelle, notamment la manière dont s'effectue, dans l'interaction verbale, la transmission des savoirs académiques ou scientifiques (Bouchard et Parpette, 2008). Le projet Scientext s'est centré quant à lui de manière exclusive sur l'analyse des discours scientifiques écrits, en prenant en compte la diversité de ses genres, articles scientifiques, communications écrites, mais aussi thèses et habilitations à diriger des recherches (HDR). Il s'est donné pour mission de constituer un corpus permettant des analyses linguistiques sur les écrits scientifiques intéressant les linguistes, mais aussi les épistémologues, les sociologues des sciences, ou encore les spécialistes de l'extraction d'information qui cherchent à identifier des passages spécifiques, comme les références

4. Voir cependant Grossmann (dir.) (2010).

« L'écrit scientifique : du lexique au discours », Agnès Tutin et Francis Grossmann (dir.)
ISBN 978-2-7535-2846-8 Presses universitaires de Rennes, 2014, www.pur-editions.fr

à autrui (par exemple, Siddhartan et Teufel, 2007), aspects qui représentent un véritable enjeu pour la veille technologique. Pour faciliter le recueil de données pertinentes, nous avons également élaboré différents outils de requête autour de deux thèmes principaux : le positionnement, dans le sens qui vient d'être explicité, et le raisonnement, en considérant les deux aspects comme complémentaires : à travers le positionnement, l'auteur s'inscrit comme sujet par rapport à ses devanciers, à ses contemporains ; il définit sa spécificité, ses choix. L'étude du raisonnement permet de retracer son cheminement intellectuel, ce sur quoi il s'appuie et les déductions qu'il opère.

La méthodologie employée s'inscrit pleinement dans la linguistique de corpus, puisqu'elle 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 al.*, 2011), comme on le verra dans la section suivante. L'analyse du discours, la lexicologie et l'approche énonciative sont aussi convoquées pour l'analyse linguistique des marques du positionnement et du raisonnement.

Pour échapper à un émiettement qui risque de conduire à la dispersion des analyses, nous avons fait le choix d'intégrer l'approche linguistique à une approche rhétorique du texte, ce terme étant entendu au sens large, c'est-à-dire prenant en compte les dispositifs argumentatifs liés aux différents genres concernés, les routines phraséologiques mais aussi les combinaisons libres des mots du discours, tout comme certains des procédés traditionnellement étudiés dans la rhétorique des figures. Il faudrait ajouter la dimension sémiologique et scripturale, encore trop souvent absente des analyses linguistiques (voir cependant Jacques, 2005), et cependant importante si l'on veut prendre en compte les spécificités de l'écrit scientifique, notamment à travers le rôle des schémas, des figures, qui impliquent des formes de circulation textuelle typiques de l'écrit : cet aspect ne faisait pas directement partie du projet, centré sur la phraséologie, mais a pu être abordé indirectement, à travers la question des marqueurs de renvois du type *voir* et cf. (Grossmann et Tutin, 2010a ; 2010b), ou dans le cadre d'un numéro de la revue *Revue d'Anthropologie des Connaissances* en particulier à travers les contributions d'Allamel-Raffin (2010) et Rabatel (2010). Marie-Paule Jacques, dans ce volume, s'intéresse de son côté au titrage et la structuration textuelle. Une telle approche intégrative ne prend sens que si l'on relie la question des genres de discours à celle des réseaux sociaux et des communautés scientifiques. Celle des identités disciplinaires – avec toutes les ambiguïtés charriées par le terme – reste un aspect qui ne peut pas non plus être négligé, étant donné les différences parfois marquées d'une discipline, ou d'un champ scientifique à l'autre.

Au vu de ce qui vient d'être rappelé, un élément clé du projet consistait à mieux articuler le plan lexical et phraséologique, au niveau énonciatif mis en jeu par la réflexion sur le positionnement. L'étude de la phraséologie de l'écrit scientifique, facilitée aujourd'hui par le recours à de larges corpus vise à dégager les spécificités du lexique scientifique,

« L'écrit scientifique : du lexique au discours », Agnès Tutin et Francis Grossmann (dir.)
ISBN 978-2-7535-2846-8 Presses universitaires de Rennes, 2014, www.pur-editions.fr

qu'il s'agisse du lexique propre à une discipline ou du lexique « transdisciplinaire » (Tutin, 2007b), c'est-à-dire du fonds commun lexical propre à la démonstration scientifique : *faire une hypothèse, montrer que, admettre le postulat selon lequel*, etc., dont l'apparition au sein de disciplines multiples ne doit pas masquer les différences d'emploi. Cet ensemble phraséologique comporte d'autres types de marques, notamment les marqueurs de métadiscours (Hyland, 2005) qui incluent l'ensemble des signaux visant à guider le lecteur et signalant le point de vue de l'auteur (évaluatifs, modalisateurs, verbes d'opinion, etc.). Des expressions comme *méthode prometteuse, résultats décevants, approche inadaptée* (Tutin, 2010c) renvoient à la fois à un discours sur l'objet scientifique (de type métascientifique) et au positionnement de l'auteur (métadiscours). Leur interprétation est tributaire du contexte énonciatif : ainsi, une collocation usuelle telle que *faire une hypothèse* peut être prêtée à autrui, assumée en son nom propre, ou au nom d'une équipe de recherche, et prend alors des valeurs argumentatives différentes.

1. 2. La question de l'unicité du texte scientifique

La première difficulté concerne l'objet même de l'investigation : la notion de discours ou texte scientifique semble reposer sur le postulat d'une forme d'unicité du texte scientifique. Or, les partitions institutionnalisées induisent d'emblée des dichotomies telles que sciences exactes *vs* sciences humaines, sciences fondamentales *vs* sciences appliquées. Faut-il, par exemple, proposer un traitement spécifique pour l'écriture scientifique dans les sciences humaines et sociales ? Quoi de commun entre un article d'ethnologie, qui peut adopter la forme d'un récit, et un article de physique, qui se plie au format IMRaD[5] ? Il peut sembler plus efficace de comparer des disciplines appartenant à une même famille d'écrits scientifiques, plutôt que de tenter de rapprocher artificiellement des écrits que tout semble opposer. La spécification de sous-genres, comme l'*article expérimental* (Bazerman, 1988), directement reliés à la forme concrète que revêt l'activité scientifique plaiderait également en faveur d'une différenciation nette. Bien que ces remarques soient fondées, un de nos postulats a été de considérer qu'il y a place également pour une approche plus décloisonnée : plaident en ce sens aussi bien l'évolution actuelle des formes du texte scientifique que les considérations – souvent formulées – portant sur les règles générales du raisonnement scientifique. Ce constat nous a conduits, pour la définition des outils de requêtes du corpus que pour son analyse, à nous intéresser également à des schémas rhétoriques et au lexique « transdisciplinaire » propre à toute communication scientifique, sans en conclure pour autant que l'utilisation de ces formes communes traduisent toujours une identité de sens ou d'emploi.

Traditionnellement, les études de la science se sont prioritairement intéressées aux sciences exactes, parce qu'il est entendu que celles-ci incarnent, de la manière la plus typique, les procédés de démonstration et de preuve mis en œuvre dans les démarches que l'on cherche à analyser. Or, il apparaît également intéressant d'étudier les schémas

5. Acronyme pour **I**ntroduction, **M**ethods, **R**esults **a**nd **D**iscussion, format standardisé pour l'article expérimental.

« L'écrit scientifique : du lexique au discours », Agnès Tutin et Francis Grossmann (dir.)
ISBN 978-2-7535-2846-8 Presses universitaires de Rennes, 2014, www.pur-editions.fr

de démonstration et les formes d'argumentation mobilisés dans les sciences humaines et sociales. Force est d'ailleurs de constater que le terme « démonstration » renvoie à des pratiques et des objets eux-mêmes très différents (Rosental, 2003), y compris au sein des sciences déductives. Ces remarques nous ont conduits, pour la constitution du corpus à refuser d'autonomiser des domaines qui seraient présentés a priori comme radicalement différents, tant du point de vue des démarches que de l'activité d'écriture. Rapprocher sans confondre, montrer les évolutions convergentes sans pour autant nier les différences fondamentales liées aux objets et aux cultures, fonde la démarche que nous avons entreprise. Entériner une approche strictement autonomiste, enfermant l'écriture de « la » Science dans des canons fixés une fois pour toutes, ignorant les influences internes ou externes aux genres scientifiques aurait conduit à commettre une erreur de méthode et se priver d'instruments d'analyse pertinents.

Dans le cadre du projet, nous avons toutefois souhaité fonder nos observations sur des faits linguistiques authentiques. Cela nous a amenés à constituer un corpus diversifié d'écrits scientifiques, relevant de trois familles de disciplines (sciences humaines et sociales, sciences expérimentales, sciences appliquées) (cf. en 2.). Pour les raisons expliquées plus haut, la plupart de nos études ont porté sur des comparaisons mettant en jeu des disciplines proches, principalement en sciences humaines. Cela nous a d'ailleurs conduits à observer des disparités tout à fait révélatrices de l'hétérogénéité de cette famille de disciplines (cf. section suivante).

1. 3. Auctorialité et positionnement de l'auteur

Un de nos objectifs était d'approfondir l'étude de la présence auctoriale et du positionnement de l'auteur dans son texte. Les écrits scientifiques sont souvent considérés comme un genre dépersonnalisé, avec un fort effacement énonciatif, où l'auteur se dissimule derrière la présentation de faits objectifs et des modalités de raisonnement partagées par la communauté scientifique. Les travaux accomplis sur ce sujet dans les dernières années (par exemple, Swales, 1990 ; Hyland, 2005 ; Fløttum *et al.,* 2006a ; Rinck, 2006) montrent cependant que ce constat est à nuancer, en tout cas dans certaines disciplines, et que l'écrit scientifique est véritablement un texte argumentatif où la dimension rhétorique est fortement présente. Fløttum *et al.* (2006a) en examinant un corpus varié d'écrits scientifiques en sciences humaines (linguistique), sciences sociales (économie) et sciences expérimentales (médecine) ont ainsi mis en évidence, à travers l'étude de plusieurs marques linguistiques énonciatives, une importante présence de l'auteur en sciences humaines et en sciences sociales. Ce constat rejoint nos propres résultats sur plusieurs aspects du positionnement de l'auteur.

L'auteur scientifique écrit avant tout pour ses pairs, en se situant au sein de traditions rhétoriques et disciplinaires particulières. Cependant, si la notion d'auteur a nourri des travaux féconds en sciences humaines depuis les travaux fondateurs de M. Foucault (1969), elle pose des problèmes spécifiques en raison de son statut épistémologique

« L'écrit scientifique : du lexique au discours », Agnès Tutin et Francis Grossmann (dir.)
ISBN 978-2-7535-2846-8 Presses universitaires de Rennes, 2014, www.pur-editions.fr

même et du feuilleté définitoire qu'elle implique, à travers ses dimensions communicatives, éditoriales et juridiques. Pour le linguiste, elle ne va pas sans poser des difficultés, puisqu'un des efforts des linguistiques de l'énoncé et du discours a justement consisté à distinguer nettement les places occupées dans un énoncé – telles qu'en rendent compte par exemple les indices personnels – de l'être juridique, textuellement responsable, qu'est l'auteur. Cette notion renvoie en effet à un statut social, individuel, ou empirique, situé a priori hors du champ d'investigation de la linguistique. C'est pour cette raison que toute appréhension naïve, consistant par exemple à partir d'un décompte des pronoms personnels à induire telle ou telle « figure » d'auteur, est inévitablement vouée à l'échec, parce qu'elle ne prend pas en compte l'autonomie, au moins relative du plan linguistique, et la complexité même des différents plans mobilisés.

Faut-il alors renoncer à établir toute relation entre la caractérisation des styles d'auctorialité scientifiques, le lexique utilisé, et l'appareil formel de l'énonciation ? Il nous a semblé possible d'établir cette jonction, au travers d'une onomasiologie du positionnement prenant en compte le système complexe des relations entre les deux plans.

Il est possible alors, à condition de bien distinguer le plan linguistique du plan éditorial, d'intégrer les aspects énonciatifs à la question de l'auctorialité scientifique. Cette démarche intégrative conduit à des notions bien connues en linguistique du discours et en sociolinguistique : identité, attitude, ethos, image de soi, posture, position, autorité, etc. Les pronoms personnels (*on*, *nous*, et *je*) ont fait l'objet de travaux assez nombreux, par exemple lorsqu'il s'agit d'opposer les usages dans des langues différentes, ou dans des disciplines différentes (cf. Breivegga, Dahl. et Fløttum, 2002 ; Fløttum, 2005 ; Fløttum *et al.,* 2006a). C'est un terrain qui se révèle très complexe, d'abord en raison de la polysémie de ces marques (cf. l'ambivalence du *on* français qui peut avoir une valeur intégratrice équivalent à un *nous* et le *on* à valeur impersonnelle) et qui suppose, comme nous l'avons signalé plus haut, de prendre garde à une vision réductrice qui assimilerait de manière directe l'emploi de telle ou telle marque à telle ou telle forme de positionnement.

L'étude du mode de référencement des sources dans l'écrit scientifique ouvre sur une problématique très spécifique, connue en linguistique sous le nom d'évidentialité, qui correspond au fait que les langues naturelles marquent différemment la source ou plutôt le mode de recueil de l'information (certaines d'entre elles spécifient même, par la morphologie, la nature de l'information obtenue – par inférence, par ouï-dire, ou par une information visuelle). En dehors de ce sens linguistique *stricto sensu*, on peut élargir la perspective comme l'a fait Chafe (1986) à la dimension discursive, particulièrement dans les genres scientifiques. Cet élargissement semble intéressant dans le cas de l'écrit scientifique, qui doit spécifier la nature des informations qu'il transmet, et des savoirs qu'il construit. Il fait appel plus spécifiquement à un certain type de garanties, extralinguistiques, mais introduites linguistiquement : références à des sources publiées, garanties empiriques (expérimentations, statistiques, tests de grammaticalité en linguistique, etc.).

Au-delà des travaux sur le lexique évaluatif, des verbes de positionnement et de la conformité/non-conformité aux attentes (Boch *et al.,* 2007 ; Cavalla et Tutin, à paraître :

« L'écrit scientifique : du lexique au discours », Agnès Tutin et Francis Grossmann (dir.)
ISBN 978-2-7535-2846-8 Presses universitaires de Rennes, 2014, www.pur-editions.fr

Tutin, 2010a, 2010c), nous nous sommes également intéressés dans le cadre du projet Scientext à la forme que revêt le cadre théorique dans lequel l'auteur s'inscrit explicitement. Il peut s'agir de la filiation intellectuelle, terme qui recouvre l'approche, les idées, voire la terminologie dont un auteur s'inspire, ou le cadre théorique dans lequel il s'inscrit explicitement (Rinck *et al.*, 2007 ; Garcia, 2008 ; Grossmann *et al.*, 2009 ; Florez, ce volume).

1. 4. Raisonnement et preuve dans l'écrit scientifique

Cet aspect du projet – le plus directement lexicologique – a été principalement pris en charge par l'équipe LiCoRN (Lorient), dont les objectifs spécifiques ont été de mettre en évidence la manière dont le raisonnement scientifique, impliqué dans les énoncés des textes scientifiques issus du corpus, peut faire l'objet d'un traitement lexicographique, à travers des notions telles que celles de *résonance collocationnelle*. Si le corpus anglais, développé à Lorient, couvre plusieurs disciplines académiques, le dictionnaire, en cours de réalisation, s'appuie plus spécifiquement sur le corpus BMC (voir la composition des corpus dans la deuxième section). Il s'agit par exemple, en étudiant les réseaux collocationnels développés autour de certains lexèmes spécifiques du domaine scientifique (Williams, 2007 ; Williams, 2010a ; Williams et Millon, ce volume), par exemple *probe* (substantif et verbe), de mieux comprendre la manière dont est utilisé le lexique propre au raisonnement ou à la démonstration, tout en observant les passages du sens général au sens spécialisé, dans les différents genres observés. Un autre angle de vue sur la preuve a été effectué à travers la question des renvois aux sources, permise par des marqueurs lexicaux tels que *voir* ou scripturaux tels que cf. (nous y revenons ci-dessous). Enfin, des travaux menés sur les verbes causatifs et leur lien au raisonnement montrent un fonctionnement spécifique des disciplines (linguistique et psychologie, biologie et médecine, électronique et mécanique) (Bak et Novakova, ce volume).

2. Les ressources textuelles élaborées dans le cadre de Scientext

Le projet Scientext avait un double objectif : d'une part, comme développé dans la section 1, il voulait effectuer une étude théorique des écrits scientifiques, en particulier sur les thèmes du positionnement et du raisonnement, à partir des marqueurs lexicaux, grammaticaux et énonciatifs ; d'autre part, il comportait un volet ingénierique avec la constitution de corpus d'écrits scientifiques librement disponibles et consultables en ligne. Nous aborderons dans cette section les principales réalisations sur ce deuxième aspect.

L'anglais bénéficie de très nombreuses études sur l'*English for Academic Purposes*, domaine qui fait l'objet de colloques et de revues spécialisées[6], de gros corpus écrits et

6. Par exemple, le *Journal of English for Academic Purposes.*

« L'écrit scientifique : du lexique au discours », Agnès Tutin et Francis Grossmann (dir.)
ISBN 978-2-7535-2846-8 Presses universitaires de Rennes, 2014, www.pur-editions.fr

oraux largement diffusés, par exemple le *Michigan Corpus of Academic Spoken English* (MICASE[7]), le *British Academic Written English* (BAWE[8]) ou les écrits scientifique du *British National Corpus* (BNC[9]). Les travaux menés à l'heure actuelle dans ces domaines sont très largement basés sur l'étude de corpus électroniques, qu'il s'agisse du discours oral ou écrit. Pour le français, à notre connaissance, si certains corpus ont été élaborés (corpus du projet KIAP[10], Fløttum *et al.* 2006a), corpus de sciences du langage et lettres constitué par Fanny Rinck, 2006), ceux-ci ne comportent pas d'annotation linguistique et ne sont pas diffusés dans la communauté scientifique. Dans le cadre du projet Scientext, nous souhaitions élaborer un ensemble de corpus, dont un corpus représentatif des écrits scientifiques du français, qui servirait d'appui aux études linguistiques envisagées, ainsi que des outils logiciels « conviviaux » permettant aux linguistes d'exploiter des corpus annotés syntaxiquement et structurellement. Une large partie du corpus français a été rendue disponible dans la communauté scientifique à des fins de recherche et tous les corpus élaborés dans le cadre du projet sont librement consultables en ligne (sur [http://scientext.msh-alpes.fr]).

2. 1. Les corpus Scientext

Dans le cadre du projet, regroupant des anglicistes et des spécialistes du français, plusieurs corpus ont été élaborés, répondant chacun à des besoins et des études linguistiques différents. Deux corpus d'écrits scientifiques en français et en anglais ont été constitués, ainsi qu'un corpus d'écrits universitaires en anglais langue étrangère et un corpus original d'évaluations de propositions de communications.

Le corpus d'écrits scientifiques du français, élaboré par le LIDILEM, comporte 4,8 millions de mots et a été conçu pour être représentatif des différents genres et disciplines scientifiques. Il était bien entendu impossible d'inclure dans le présent projet la totalité ou la quasi-totalité des disciplines représentées, par exemple par les différentes sections du CNRS ou des sections du CNU, mais nous avons 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 (la linguistique, la psychologie, les sciences de l'éducation et dans une certaine mesure, le traitement automatique des langues), les sciences expérimentales (biologie, médecine) et les sciences appliquées ou sciences pour l'ingénieur (électronique, mécanique), les frontières entre ces familles n'étant bien entendu pas étanches. Les sous-genres sélectionnés intègrent des articles de recherche, des communications écrites, des thèses de doctorat et des mémoires d'habilitation à diriger les recherches. Le corpus public, consultable en ligne, compte 4,8 millions de mots. Un corpus plus large (8,4 M de mots), intégrant davantage de disciplines, en particulier en sciences humaines,

7. [http://micase.elicorpora.info/]
8. [http://www2.warwick.ac.uk/fac/soc/al/research/collect/bawe/]
9. [http://www.natcorp.ox.ac.uk/]
10. KIAP : Kulturell Identitet i Akademisk Prosa (Identité culturelle dans les écrits scientifiques).

« L'écrit scientifique : du lexique au discours », Agnès Tutin et Francis Grossmann (dir.)
ISBN 978-2-7535-2846-8 Presses universitaires de Rennes, 2014, www.pur-editions.fr

est disponible exclusivement en intranet pour les membres du projet, pour des raisons de droits. Le corpus a été annoté sur le plan morphologique et syntaxique et sur le plan structural selon le format XML en suivant les recommandations de la Text Encoding Initiative (cf. plus bas). Un sous-ensemble du corpus public est librement disponible à des fins de recherche (corpus de 3,9 M de mots). Le corpus français a été exploré dans le cadre de plusieurs études linguistiques au LIDILEM (cf. ci-dessous). Il est à nouveau utilisé dans le cadre d'un nouveau projet ANR-Contint piloté par l'ATILF, auquel le LIDILEM participe.

Le corpus d'écrits scientifiques anglais a été élaboré par l'équipe LiCorn de l'Université de Bretagne Sud (Geoffrey Williams, Chrystel Millon). Les textes proviennent de la maison d'édition indépendante BioMed Central et portent exclusivement sur la biologie et la médecine et le corpus d'origine est librement téléchargeable [11]. Ce corpus de 13,9 M de mots se compose essentiellement d'articles de recherche et de communications écrites (90 % du corpus) mais comporte aussi des rapports, enquêtes, bases de données, articles courts, etc. Comme le corpus français, le corpus a été annoté du point de vue structural (suivant les recommandations de la TEI) et annoté morphologiquement et syntaxiquement. Le corpus a été exploité dans les travaux de Geoffrey Williams et Chrystel Millon, principalement sur les aspects lexicaux (cf. chapitre sur les verbes de la science dans ce volume) et fait actuellement l'objet de travaux lexicologiques au Lidilem (Laura Hartwell).

Le corpus d'écrits universitaires en anglais langue étrangère a été élaboré par le laboratoire LLS de l'Université de Savoie (John Osborne, Alice Henderson, Robert Barr). Ce corpus d'1,1 M de mots comporte des travaux d'apprenants universitaires français écrivant en anglais, principalement des étudiants de 2e et 3e année du cursus d'anglicistes apprenant à rédiger de textes argumentatifs longs (4 500 mots) qui s'appuient sur des recherches documentaires approfondies. Le corpus a été annoté du point de vue morphologique et syntaxique, ainsi qu'au niveau de la structure, mais les parties textuelles balisées restent peu nombreuses (introduction, conclusion, corps du texte, titres). Ce corpus a fait l'objet de plusieurs études linguistiques dans l'équipe LLS (cf. chapitre d'Alice Henderson dans ce volume).

Le corpus d'évaluations de propositions de communications a été conçu par le LIDILEM (Françoise Boch, Achille Falaise). Il s'agit d'un corpus original et expérimental comportant 520 commentaires évaluatifs de relecteurs pour un colloque de jeunes chercheurs en sciences du langage (Colloque CEDIL 2010). L'utilisateur peut sélectionner les commentaires selon le sous-domaine disciplinaire, selon le destinataire du commentaire (auteur ou organisateur) et selon l'évaluation attribuée (rejeté, accepté, réservé). Ce corpus a été annoté morphologiquement et syntaxiquement, mais ne comporte pas d'annotation de structure. Il a fait l'objet d'études linguistiques au sein du LIDILEM (cf. chapitre de Boch, Rinck & Nardy dans ce volume).

Comme indiqué précédemment, le corpus a été annoté à plusieurs niveaux, en suivant les recommandations en vigueur (XML-TEI, principalement) dans la communauté de la

11. [http://www.biomedcentral.com/about/datamining]

« L'écrit scientifique : du lexique au discours », Agnès Tutin et Francis Grossmann (dir.)
ISBN 978-2-7535-2846-8 Presses universitaires de Rennes, 2014, www.pur-editions.fr

linguistique de corpus. Chaque texte comporte ainsi un en-tête indiquant précisément à la fois la notice bibliographique du texte, mais aussi les annotations effectuées, les droits, etc.

L'annotation linguistique, qui comporte à la fois une annotation morphologique mais aussi de façon plus originale, une annotation syntaxique, a été réalisée par le logiciel Syntex élaboré par Didier Bourigault (2007)[12]. Le modèle syntaxique est celui de la syntaxe de dépendance, les mots étant rattachés à d'autres mots par des relations syntaxiques de type Sujet, Objet, Épithète, etc. La figure ci-dessous fournit une représentation graphique de l'analyse de la phrase : *chaque puce comporte huit spots permettant d'analyser huit échantillons.*

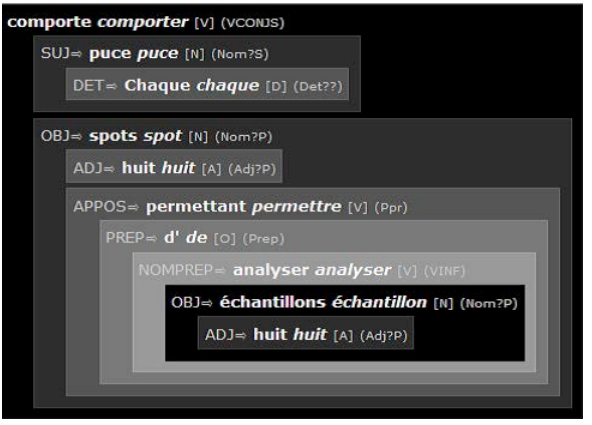

Figure 1 : Représentation de l'analyse syntaxique de *chaque puce comporte huit spots permettant d'analyser huit échantillons*

L'analyse morphologique indique la catégorie grammaticale du mot, le lemme (la forme conventionnelle, par exemple l'infinitif pour le verbe), quelques sous-catégories ou marques de flexion. L'analyse syntaxique indique les fonctions syntaxiques et les liens entre les mots. La tête de la phrase est ainsi le verbe *comporter* auquel sont rattachés deux mots, *puce* et *spots*, par des relations de Suj(et) et d'Obj(et). D'autres mots leur sont ensuite rattachés. L'analyse syntaxique présente de nombreux intérêts : (1) elle permet de faire des recherches dans les textes avec des patrons plus abstraits (par exemple, « je cherche tous les verbes qui ont *hypothèse* comme objet direct ») ; (2) elle permet d'extraire des éléments qui sont éloignés dans la phrase, séparés par des incises par exemple. (3) elle s'affranchit de la linéarité : l'objet direct peut ainsi être placé avant ou après le verbe. Ces avantages ont toutefois un revers : proposer un outil simple d'exploitation de ces structures est un défi, relevé avec succès par Achille Falaise et Olivier Kraif (cf. chapitre d'Achille Falaise dans ce volume).

12. Que nous remercions vivement ici.

« L'écrit scientifique : du lexique au discours », Agnès Tutin et Francis Grossmann (dir.)
ISBN 978-2-7535-2846-8 Presses universitaires de Rennes, 2014, www.pur-editions.fr

Pour finir, une annotation structurale a été également proposée. Cette annotation, qui suivait les recommandations de la Text Encoding Initiative, permettant de distinguer les principales parties textuelles de l'écrit scientifique (introduction, développement, conclusion, résumé, notes de bas de page, bibliographie, annexes). Elle a permis des études linguistiques ciblées, sur les spécificités rhétoriques des introductions ou conclusions par exemple, ou d'écarter dans les statistiques lexicales les parties contenues dans la bibliographie. En outre, autant que faire se peut, les styles d'origine (italiques, gras, mise en page) ont été conservés de façon à permettre à plus long terme des études sur la structure textuelle.

Le tableau 1 résume les différents types de corpus élaborés dans le cadre du projet

	Écrits scientifiques du français	**Écrits scientifiques de l'anglais**	**Écrits universitaires argumentatifs en anglais langue étrangère**	**Évaluations de propositions de communications**
Conception et annotations	LIDILEM	Équipe LIcorn	LLS	LIDILEM
Domaine ou discipline	Pour le corpus public, linguistique, psychologie, sciences de l'éducation, électronique, informatique, biologie.	Médecine et biologie	Sujets de sociolinguistique et société	Linguistique (Colloque jeunes chercheurs en linguistique)
Type d'écrit (genre)	Articles de recherche, communications écrites, thèse de doctorat, mémoires d'HDR	Articles, communications, mais aussi des rapports, des synthèses…	Écrits argumentatifs d'étudiants anglicistes de 2e et 3e année	Évaluations de propositions de communications
Volume	4,8 M de mots pour le corpus public (en ligne) 8,4 M de mots pour le corpus interne à l'équipe. 3,9 M de mots pour le corpus diffusé.	13,8 M de mots	1,1 M de mots	35 000 mots
Annotations	– Syntaxiques et morphosyntaxiques – Annotation structurale détaillée des parties textuelles	– Syntaxiques et morphosyntaxiques – Annotation structurale des parties textuelles	– Syntaxiques et morphosyntaxiques – Annotations structurales	– Syntaxiques et morphosyntaxiques
Disponibilité	Corpus public Interrogeable en ligne [13] Un sous-ensemble du corpus librement diffusé.	Interrogeable en ligne Corpus disponible sur le site de Biomed Central	Interrogeable en ligne Non diffusé	Interrogeable en ligne Non diffusé

13. Le corpus « privé » (du fait de droits indisponibles) n'est disponible qu'en intranet.

« L'écrit scientifique : du lexique au discours », Agnès Tutin et Francis Grossmann (dir.)
ISBN 978-2-7535-2846-8 Presses universitaires de Rennes, 2014, www.pur-editions.fr

2. 2. Interrogation des corpus de Scientext : le logiciel Concquest

Les corpus de Scientext sont interrogeables en ligne à l'aide d'un outil logiciel développé par Achille Falaise (cf. chapitre dans ce volume), basé sur un moteur de recherche de textes annotés linguistiquement, ConcQuest, élaboré par Olivier Kraif (2008).

Dès le début du projet, il nous est apparu essentiel de réfléchir à l'ergonomie de cet outil qui conditionnait l'utilisation du corpus par les linguistes, et tout au long du processus d'élaboration, les utilisateurs ont été associés au développement de l'outil (Falaise *et al.*, 2011). Les fonctionnalités du logiciel étant décrites en détail dans le chapitre d'Achille Falaise dans ce volume, nous ne passerons pas ici en revue toutes ses fonctionnalités, mais en soulignerons simplement les principales caractéristiques. L'outil a été développé pour être facilement utilisable par des non spécialistes, et l'interface graphique utilise abondamment des listes à choix et des ascenseurs.

– Choix des textes :

À la façon de Frantext, l'utilisateur peut sélectionner son corpus d'après différents critères (comme les disciplines, les parties textuelles, les types textuels), écarter certains textes jugés non pertinents et mémoriser le corpus sélectionné pour une utilisation ultérieure.

– Recherche dans les textes :

Plusieurs modes de recherche dans les textes sont proposés : a) un mode guidé et simple, ne requérant pas de compétence spécifique ; b) un mode sémantique, où des requêtes préétablies d'après un thème, par exemple, le point de vue de l'auteur, ont été enregistrées (cf. chapitres de Falaise et Tutin dans ce volume pour des exemples) ; c) un mode avancé, où l'utilisateur doit maîtriser un langage de requête assez complexe mais qui permet de formuler des requêtes plus spécifiques. Les recherches peuvent porter sur les formes, les catégories syntaxiques, les lemmes, et les relations syntaxiques entre les mots. Les statistiques d'utilisation du système (cf. chapitre de Falaise dans ce volume) montrent que le mode simple et le mode sémantique sont très largement préférés par les internautes utilisateurs, ce qui conforte le choix fait dans notre projet de mettre en avant les aspects ergonomiques dans l'utilisation de corpus annotés syntaxiquement et structurellement.

– Affichage des résultats :

L'outil affiche des concordances KWIC, qui peuvent être élargies à la demande. L'utilisateur peut sélectionner un sous-ensemble de résultats et l'exporter dans différents formats. Des statistiques simples permettent de générer des fréquences et des répartitions par partie textuelle, discipline ou genre textuel. Ces derniers éléments permettent des comparaisons intéressantes entre les différents types d'écrits.

À l'heure actuelle, l'outil est largement utilisé au sein de l'équipe et à l'extérieur. Il est globalement satisfaisant, même si le temps de réponse pour certaines requêtes complexes doit encore être amélioré. Nous sommes persuadés que l'utilisation de corpus avec des annotations de haut niveau ne se popularisera qu'avec le développement d'outils convi-

« L'écrit scientifique : du lexique au discours », Agnès Tutin et Francis Grossmann (dir.)
ISBN 978-2-7535-2846-8 Presses universitaires de Rennes, 2014, www.pur-editions.fr

viaux conçus avec les utilisateurs et nous espérons avoir contribué dans le cadre du projet Scientext à cette réflexion.

3. Les contributions du volume

Les contributions du présent volume rendent compte de la diversité des études réalisées dans le cadre du projet Scientext. Elles portent sur tous les types de corpus, qu'il s'agisse du français ou de l'anglais, du corpus scientifique ou du corpus d'apprenants. Le corpus plus expérimental d'évaluations de communications a aussi fait ici l'objet d'une contribution. En outre, différentes approches linguistiques sont proposées, des études lexicales et phraséologiques à l'analyse de discours. Les aspects théoriques sont développés à côté d'approches plus applicatives, qu'il s'agisse de la didactique ou du traitement automatique des langues.

La première série d'écrits présente les outils théoriques et méthodologiques de l'étude du positionnement et du raisonnement de l'auteur dans les écrits scientifiques.

Agnès Tutin dans la continuité de travaux sur le lexique scientifique transdisciplinaire (Tutin, 2007a ; Tutin, 2010b) aborde la question des séquences polylexicales propres aux écrits scientifiques, une entrée privilégiée dans l'approche linguistique du projet, mais qui constitue une catégorie hétérogène dont il convient de proposer une modélisation adaptée. Elle propose une typologie fonctionnelle, un peu à la façon de Granger et Paquot (2008) en distinguant les principales fonctions remplies par ces expressions : référentielle, discursive, interpersonnelle, rhétorique. Elle met ensuite l'accent sur les routines sémantico-rhétoriques des écrits scientifiques pour lesquelles elle propose un cadre de description s'inspirant de la sémantique des cadres de Fillmore.

Williams & Millon approfondissent ici des travaux abordés dans Williams & Millon (2009) et Williams & Millon (2010). Ils présentent la structure d'un dictionnaire électronique général d'encodage anglais des sciences destiné aux locuteurs non natifs, et particulièrement aux scientifiques. Ce dictionnaire a pour objectif d'aider les locuteurs non natifs à passer du général au spécifique, et inversement du spécifique au général. Il s'agit d'un dictionnaire phraséologique avec une catégorisation conceptuelle des verbes, où les entrées lexicales se composent des définitions des différents sens des mots, mais également des collocations et des patrons lexico-syntaxiques typiques. La compilation du dictionnaire est entièrement guidée par les données d'un corpus.

Magda Florez prolonge les travaux effectués dans le cadre de Scientext autour du cadre théorique et de la référence à autrui (Boch *et al.*, 2009 ; Grossmann *et al.*, 2009). Elle s'intéresse à l'étude linguistique du positionnement à travers la « citation positionnée », c'est-à-dire les références aux pairs vis-à-vis desquelles l'auteur se positionne explicitement dans l'écrit scientifique, qu'il y adhère ou qu'il s'en démarque. Après un état de l'art sur les différents types d'études sur la référence à autrui, elle présente les structures sémantiques et syntaxiques de la citation. Son étude de cas sur les écrits de psycholo-

« L'écrit scientifique : du lexique au discours », Agnès Tutin et Francis Grossmann (dir.)
ISBN 978-2-7535-2846-8 Presses universitaires de Rennes, 2014, www.pur-editions.fr

gie, sciences de l'éducation et linguistique montre des différences remarquables dans les pratiques citationnelles, plus marquées dans les articles de recherche que dans les thèses.

Francis Grossmann reprend dans ce chapitre une problématique déjà abordée autour du thème de l'évidentialité et de la construction de la preuve, principalement à travers le verbe *voir* (Grossmann et Tutin, 2010a ; Grossmann et Tutin, 2010b). Il élargit ici l'objet d'étude en abordant la question des verbes de constat dans l'écrit scientifique, à travers plusieurs lexèmes : *observer, constater, noter, remarquer, voir, s'apercevoir*. Les structures syntaxiques mais aussi sémantiques et énonciatives (le sujet inclut-il ou non le lecteur dans le constat ?) sont observées et montrent une certaine hétérogénéité de cet ensemble. Selon les verbes, les formes de constat penchent plutôt vers le factuel (*observer*) ou le versant interprétatif (*constater*), alors que toutes les formes de constat n'ont pas la même valeur informationnelle, certaines pouvant être considérées comme des constatifs plus « faibles », en particulier en cas d'incise (ex : *on l'a vu…*).

Monika Bak et Iva Novakova s'intéressent aux verbes causatifs dans les écrits scientifiques du corpus français de Scientext. Elles classent les principaux éléments du lexique verbal causatif et en étudient la combinatoire lexicale et syntaxique. L'étude montre l'existence d'un lexique verbal causatif varié et largement transversal. L'observation des répartitions montre également que certaines disciplines comme la psychologie cognitive et sociale se rapprochent sur ce point de certaines disciplines expérimentales comme la biologie ou la médecine. En outre, les formes causatives observées révèlent souvent une présence énonciative forte.

La première partie s'achève avec la présentation par Achille Falaise des outils (corpus, logiciel d'exploitation, site) développés dans le cadre de Scientext. Les corpus constitués et les annotations réalisées sont présentés ainsi que l'outil d'exploitation développé pour interroger les corpus. Des exemples d'exploitation sont présentés.

La deuxième partie aborde les aspects didactiques. Alice Henderson propose une analyse du sous-corpus « apprenants » de Scientext : 300 écrits universitaires anglais d'étudiants français de premier cycle. L'analyse se focalise sur les marques de positionnement dans leurs écrits, particulièrement dans les séquences de *I* et de *we* associées à des verbes d'opinion. L'hypothèse que les étudiants de troisième année seraient plus familiarisés aux usages en cours dans l'écrit universitaire, et utiliseraient moins de marques personnelles, n'est pas confirmée. L'influence potentielle des consignes et de la langue et culture d'origine est également discutée.

Cristelle Cavalla et Mathieu Loiseau discutent ensuite une utilisation du corpus Scientext pour la didactique du Français Langue Étrangère. Le corpus est ainsi détourné de sa fonction première pour une utilisation didactique. Une séquence d'utilisation, centrée sur la phraséologie scientifique, est présentée. La démarche proposée permet de familiariser les apprenants avec les collocations « transdisciplinaires », qui sont extraites du corpus de manière à les aider dans leurs activités rédactionnelles ; ce système d'aide à la rédaction s'appuie sur un système de « fonctions rhétoriques » (filiation, démarcation, etc.) du type de celui qui est présenté dans Scientext.

« L'écrit scientifique : du lexique au discours », Agnès Tutin et Francis Grossmann (dir.)
ISBN 978-2-7535-2846-8 Presses universitaires de Rennes, 2014, www.pur-editions.fr

La troisième et dernière partie aborde les aspects discursifs et textuels. Françoise Boch, Fanny Rinck et Aurélie Nardy s'intéressent à un type d'écrit scientifique encore peu étudié, les évaluations de proposition de communication. Elles analysent la rhétorique de l'évaluation, et en particulier comment l'expert s'adresse à l'auteur, et si ces marques d'adresses varient suivant le diagnostic qu'il pose sur la proposition de communication. Elles montrent en particulier que l'utilisation du pronom de première et de deuxième personne varie fortement en fonction de l'orientation de l'évaluation (plutôt *vous* pour les évaluations négatives, *je* pour les évaluations positives). En outre, l'étude révèle que les termes d'évaluation utilisés sont fortement routinisés.

Enfin, Marie-Paule Jacques étudie, à partir d'un corpus d'une trentaine d'articles de trois disciplines de Sciences Humaines et Sociales, les modes de structuration textuelle, tels qu'ils apparaissent à travers les intertitres. Son analyse met en évidence la diversité de ces modes de structuration et le fait que la structure de l'article est modelée par les étapes et opérations de la recherche, mais peut aussi être conditionnée par la dimension argumentative des textes.

« L'écrit scientifique : du lexique au discours », Agnès Tutin et Francis Grossmann (dir.)
ISBN 978-2-7535-2846-8 Presses universitaires de Rennes, 2014, www.pur-editions.fr