Etude lexicographique et discursive des collocations en vue de leur intégration dans une base de données terminologiques

Mojca Pecman, Université Paris Diderot – Paris 7

RESUME

L'article présente une étude discursive des collocations, dont la visée est d'améliorer le traitement lexicographique de ce type d'unités linguistiques dans une base de données terminologiques en ligne, appelée base ARTES (Aide à la Rédaction de TExtes Scientifiques). Cette base sert à la création de ressources lexicales destinées aux traducteurs spécialisés et aux scientifiques souhaitant rédiger une communication en langue seconde. L'enregistrement des collocations dans la base permet de fournir des informations précieuses sur l'utilisation des termes en langues de spécialité. Une telle entreprise soulève nombre de questions lexicographiques liées à la problématique du traitement différentiel des collocations spécifiques aux domaines et de celles qui sont spécifiques aux genres. Nous procédons ainsi à une analyse discursive des phénomènes collocationnels les plus saillants dans les articles scientifiques portant sur les domaines des sciences de la terre, de l'environnement et de l'écologie. Un corpus a été créé en dégageant les différentes sections des articles afin de permettre une analyse de la distribution des collocations à travers les textes. Cette étude ouvre la voie à une interprétation des collocations en fonction de leurs rôles spécifiques au niveau du genre, du discours, du domaine et du texte.

ABSTRACT

In this paper we propose a discursive analysis of collocations with a view to improving a lexicographical approach to this type of linguistic items within an online terminological database, known as the ARTES database (Aide à la Redaction de TExtes Scientifiques / Dictionary-assisted writing tool for scientific communication). The target users of this database are specialised translators and scientists who need to write academic papers in their second language. Collocations can provide useful information about the way terms are used in Languages for Special Purposes. The paper discusses a number of lexicographical issues raised by such a project: namely the difficulty in distinguishing domain-specific collocations from genre-specific collocations when building up phraseological resources. In order to shed light on this problem, we conduct a discursive analysis of the most salient collocational phenomena observed in a corpus of scientific articles from the field of Earth and Planetary Sciences, Environmental Sciences and Ecology. The corpus that was compiled consists of texts organised according to the structure of scientific articles, allowing for the distributional analysis of collocations across texts. This study opens up routes to an interpretation of the behaviour of collocations according to their specific roles within a genre, discourse, domain and text.

MOTS CLES

Collocations, base de données terminologiques et phraséologiques, lexicographie, analyse du discours, traduction, langues de spécialité.

KEYWORDS

Collocations, terminological and phraseological database, lexicography, discourse analysis, translation, LSPs.

1. Introduction

La présente étude porte sur la problématique de la création de ressources lexicales en vue de l'aide à la traduction et la rédaction en langue seconde dans le domaine des connaissances spécialisées. Elle s'inscrit ainsi dans la voie des recherches initiées par Henri Zinglé, dont les outils de traitement des ressources phraséologiques (Zinglé 1994, 1996, 1998), ainsi que la méthode de systématisation de la phraséologie en vue de la création de dictionnaires (Zinglé et Brobeck-Zinglé 2003), ont fortement influencé notre approche de la phraséologie scientifique. Depuis 2007, notre étude des collocations des discours spécialisés se poursuit dans le cadre du projet ARTES (Aide à la Rédaction de TExtes Scientifiques) qui a abouti récemment à la construction d'une base de données terminologiques en ligne: la base ARTES. Ce projet a la spécificité d'offrir un cadre pour l'étude des aspects lexicographiques de la création de ressources linguistiques spécialisées en s'appuyant sur l'analyse discursive du rôle des unités linguistiques dans les textes spécialisés.

La problématique que nous proposons d'explorer dans cet article est la suivante: la constitution de ressources collocationnelles en langue de spécialité peut-elle être adaptée pour refléter davantage leur fonctionnement dans le discours? Deux hypothèses de travail sont à l'origine de cette approche double des unités linguistiques, approche lexicographique et approche discursive. Notre première hypothèse est que la description et l'organisation des ressources dans une base de données fondées sur l'observation et une analyse fine du fonctionnement des unités lexicales dans le discours, permettraient de mieux rendre compte de l'usage de ces unités dans le discours. Notre deuxième hypothèse de travail est qu'une base regroupant les données terminologiques et phraséologiques permettrait d'offrir aux utilisateurs – traducteurs ou experts – des informations plus complètes pour se conformer aux conventions de communication d'un genre et d'une discipline, notamment dans une perspective de traduction ou de rédaction dans la langue seconde. En effet, parfois davantage que la terminologie, c'est souvent la phraséologie spécialisée qui est maîtrisée de façon incertaine en langue seconde.

Bien qu'à l'heure actuelle nous disposions de vastes connaissances sur ces unités complexes du lexique – connaissances accumulées grâce au regain d'intérêt général pour les collocations et la phraséologie dans les années 80 – il reste néanmoins que leur rôle discursif est encore assez peu pris en compte. Les études des collocations restent majoritairement fondées sur une analyse de corpus, généralement de grande taille, et font état des caractéristiques observées au niveau d'une langue, ou d'une langue de spécialité. Toutefois, des études commencent à démontrer que les collocations n'ont pas seulement pour fonction de marquer les profils cooccurrentiels spécifiques à une langue donnée mais également les modes d'association des termes récurrents dans une discipline et un

genre. Notre étude tentera ainsi de démontrer que la distribution de ces unités dans le texte et les variations dans leurs modes d'association constituent un facteur essentiel de la textualité: ce facteur doit être pris en compte dans la perspective d'une approche lexicographique des collocations, notamment dans le cadre de la création de ressources d'aide à la traduction et la rédaction en langue de spécialité.

Dans une première partie, la présente étude approfondira la question de la place des collocations dans l'étude des langues de spécialité, établissant un état de l'art des recherches et des nouvelles tendances en la matière. La deuxième partie est consacrée à la présentation, d'une part, du projet ARTES autour duquel s'articule la présente étude et, d'autre part, de la méthodologie développée pour observer le rôle discursif des collocations dans les textes spécialisés. La troisième partie porte sur les résultats des analyses effectuées et fait état des comportements caractéristiques des collocations dans une perspective d'interprétation discursive. La dernière partie de cette étude discute les possibilités de prise en compte de ces comportements par le modèle lexicographique d'intégration des collocations dans la base ARTES.

2. Contexte théorique

2.1. Approches dominantes dans les études des collocations

L'intérêt désormais reconnu des linguistes pour les collocations et leur rôle dans les langues de spécialité nous encouragent à proposer une analyse fine des collocations dans le discours scientifique tant de point de vue de la conception de ressources lexicales que du point de vue de leur rôle spécifique dans ce type de discours. Afin de mieux situer notre étude, une définition de la notion de collocation et une mise en perspective des approches dominantes des collocations sont nécessaires.

Les collocations ont en effet donné lieu à de nombreuses définitions cherchant notamment à les opposer aux combinaisons libres et aux expressions figées. On trouve par exemple cette opposition dans la définition de Hausmann:

> La collocation se distingue de la combinaison libre (the book is useful, das Buch ist nützlich, le livre est utile) par la combinabilité restreinte (ou affinité) des mots combinés (feuilleter un livre vs. acheter un livre). Elle se distingue des locutions (idioms, Redewendungen, par ex. monter un bateau à qn/jdn. durch den Kakao dringen/to pull sb's legs) par son non-figement et par sa transparence (1989: 1010).

Cependant, établir la limite entre les collocations, notamment nominales, et les unités lexicales composées, entraîne une autre distinction qui pose des difficultés. Une définition sommaire, néanmoins pertinente, des collocations est également possible, telle que celle de Benson (1989: 3): "arbitrary recurrent word combinations." (Pour une discussion sur la

problématique de la définition des collocations, voir par exemple Dubreuil (2008), et sur les différents types d'unités collocationnelles, voir par exemple Granger et Paquot (2008).)

Les études sur les collocations accordent une place de première importance au contexte dans le processus d'analyse et de description des unités lexicales, créant un nouveau paradigme d'observables linguistiques autour des concepts voisins, tels que celui de *collocation* exploré dans des travaux de Firth (1968), Halliday (1966) et Sinclair (1966, 1991), mais également de *pattern grammar* approfondi par Hunston et Francis (2000) et Partington (1998) ou encore de *lexical priming* introduit par Hoey (2004, 2005).

Si dans une grande partie des travaux sur les collocations et la phraséologie, l'accent est mis sur la nécessité de création de ressources collocationnelles (e.g. Hausmann 1979, Heid et Freilbott 1991, Pavel 1993, Fontenelle 1994, Benson *et al.* 1997, Meynard 1997, L'Homme et Meynard 1998, Siepmann 2006, Tutin 2007, 2008, Pecman 2008, Pecman *et al.* 2010, Kübler et Pecman 2012, etc.) et l'importance de ce type d'unité pour une bonne maîtrise de la langue seconde (e.g. Mel'čuk 1993, Howarth 1996, Granger 1998), les plus récentes avancées en la matière tentent d'évaluer leur rôle dans la construction du discours et des genres textuels (e.g. Viprey 2006, Biber *et al.* 2007, Bordet 2011). Par conséquent, après les approches didactique et lexicographique, dominantes jusqu'à présent, qui orientaient les recherches vers la création de ressources phraséologiques et les problématiques de consignation des collocations dans les dictionnaires, semble se dessiner une nouvelle voie de recherche qui vise à replacer la phraséologie dans le domaine de l'analyse du discours et de la structuration informationnelle du texte, c'est-à-dire de la construction du sens au niveau du texte. Cette nouvelle dimension présente par ailleurs l'avantage de permettre de parfaire également nos approches lexicographiques de ce type d'unité, en renforçant nos connaissances sur les collocations.

Combiner les approches lexicographique, didactique et discursive a une double conséquence. D'une part, cela permet de rapprocher la création de ressources lexicales de l'étude de la grammaire des textes. D'autre part, cela permet d'évaluer la possibilité de la mise en place d'une démarche onomasiologique dans l'analyse des lexiques spécialisés, et plus particulièrement des lexiques transdisciplinaires (Pecman 2007). L'avantage non négligeable de cette démarche est que dans la phase de création de ressources lexicales, elle permet d'envisager un double accès aux données, soit à partir de leur forme soit à partir de leur contenu sémantique.

A côté de nombreux travaux sur les collocations d'une langue de spécialité donnée (pour n'en citer que quelques-uns: Williams (1999) sur le domaine des plantes parasites, Gledhill (2000) sur le domaine de la

pharmaceutique liée à la recherche sur le cancer, Volanschi (2008) sur le domaine de la biologie de la levure), depuis une dizaine d'années, de plus en plus d'études abordent les lexiques transdisciplinaires, mettant l'accent sur la question de l'identification, de la description et de l'utilité de ce type de ressources (e.g. Coxhead 2002, Coxhead et Hirsh 2007, Pecman 2007, Tutin 2007, 2008, Simpson-Vlach et Ellis 2010). L'analyse des lexiques transdisciplinaires soulève également la question du rôle des collocations dans la construction d'un discours et d'un genre étant donné que le lexique transdisciplinaire transcende les spécificités linguistiques des domaines pour se nicher dans celles des genres et des discours.

Par-delà les domaines, les genres et les discours, la collocation est également un phénomène linguistique marquant des textes. Par conséquent, une étude sur des collocations et sur des textes, se doit de chercher un moyen de rendre compte de leur interaction.

2.2. La collocation comme élément de cohésion textuelle

Dans la partie précédente, nous avons souligné que l'approche lexicographique, orientée notamment vers la constitution de ressources pour pallier les problèmes de traduction et de maîtrise des collocations en langue seconde et/ou en langue de spécialité, a longtemps été l'approche dominante dans le contexte de l'étude des collocations. Notre étude cherche à proposer des pistes pour une prise en compte de la dimension discursive dans l'étude des collocations, tant pour aboutir à une meilleure compréhension du rôle des collocations dans la langue que pour rendre leur traitement lexicographique plus efficace. Nous pensons en effet que les méthodes lexicographiques actuelles de création de ressources collocationnelles pourraient être améliorées par une mise en perspective discursive. Cette nouvelle approche apporterait notamment des éléments d'informations intéressants pour la résolution des problèmes rédactionnels et traductionnels que le maniement des collocations peut poser au sein de l'objet clos et marqué par une finalité discursive que constitue le texte.

L'étude des collocations est généralement fondée sur l'analyse d'un corpus recueillant de grandes quantités de textes, dont les spécificités (couverture en domaine, type de document, langue, etc.) dépendent des applications visées. Pourtant, les collocations, en tant que phénomènes coocurrentiels, ne se manifestent pas uniquement à travers un corpus de textes, mais également dans une fenêtre plus restreinte, celle d'un texte. En effet, la coprésence régulière de deux unités linguistiques dans une fenêtre textuelle donnée joue un rôle important dans la cohérence globale du texte, rôle qui ne peut être efficacement perçu dans un corpus où les frontières entre textes sont rompues.

Le texte représente la dimension la plus complexe de l'aboutissement langagier. Il constitue néanmoins une unité linguistique identifiée par sa forme et ses caractéristiques, et est étudié en tant qu'ensemble clos (cf.

de Beaugrande 1980; Rastier 1987, 1989). C'est un objet suprême de la langue dont la description nécessite la prise en compte de plusieurs niveaux d'analyse: microstructure, macrostructure, suprastructure; étude de la situation d'énonciation, de l'intertextualité, pour n'en citer que quelques-uns.

Récemment, l'analyse des discours scientifiques a mené au concept de « moves » (Hyland 2004; Biber *et al.* 2007), fondé sur les travaux de Swales (1990) selon lesquels différentes portions d'un texte correspondent aux fonctions pragmatiques codées qu'elles assurent au sein d'un genre, telles que: la description du contexte théorique, la déclaration d'une intention de recherche, la précision de l'objet d'étude, la présentation des outils et des méthodes, ou encore la présentation des résultats et des conclusions auxquelles on a abouti. A ces différentes fonctions correspondraient des collocations particulières. En d'autres termes, ce sont les combinaisons stéréotypées d'items lexicaux qui contribueraient à la mise en place de ces différentes fonctions au sein d'un texte.

Les recherches sur les collocations correspondant à ces unités informationnelles codées des textes spécialisés – qu'elles soient désignées par « moves », par « mouvements rhétoriques », « fonctions rhétoriques » ou « fonctions discursives » – ont été prises en compte dans la conception de la base ARTES, notamment dans le traitement des collocations non spécifiques aux domaines (cf. section 5). Dans la terminologie de la base ARTES, nous parlons des « fonctions discursives » que nous définissons comme unités textuelles au contenu informationnel attendu, qui s'actualise dans des structures lexico-grammaticales reconnaissables au sein d'un type de discours, et, nécessairement, d'un type de genre textuel.

Plus généralement, une des propriétés centrales reconnue des textes, qu'ils soient spécialisés ou non, est la cohérence (cf. Halliday et Hasan 1976) qui repose sur divers éléments de cohésion: anaphores, isotopies, ellipses, structures à thème-rhème, etc. Ces éléments de cohésion participent, à travers le mécanisme de répétition et de progression, à la création d'un objet réticulaire complexe dont les récurrences, les échos, assurent précisément l'unité.

Il a été démontré que la cooccurrence est un élément de cohésion textuelle très fort (Halliday et Hasan 1976; Hoey 1991). Halliday et Hasan (1976: 285-6) en particulier avaient illustré le rôle des collocations continues ou discontinues binaires (e.g. *laugh…joke, blade…sharp, garden…dig, ill...doctor, try...succeed, bee...honey*), mais également des structures collocationnelles plus complexes qui forment de véritables chaînes collocationnelles (e.g. *hair...split-ends...comb...blow-dry...perm*), dans la création de la texture du texte. L'analyse présentée dans la section 4 de cet article montre que les collocations, de par leur récurrence d'apparition dans un texte, ont un pouvoir cohésif très fort.

Comme nous l'avons mentionné dans la partie précédente, la tendance actuelle est de croiser l'étude des phénomènes collocationnels avec l'analyse du discours. Ces pistes de recherche offrent un éclairage novateur sur des phénomènes cooccurrentiels complexes comme les chaînes ou réseaux collocationnels, nommés et modélisés parfois différemment selon les linguistes (cf. le concept de « cascades d'expressions » de Gledhill (2011), de « cooccurrence généralisée » de Viprey (2006) ou de « poly-cooccurrence » de Martinez (2003)). Ces réseaux construisent des structures d'équivalence, ou de résonance, relativement prévisibles mais aussi productives, qui sont comme scellées dans le texte et sur lesquelles repose l'agencement de la structure textuelle.

Enfin, plus rarement, l'analyse des phénomènes cooccurrentiels est envisagée de façon à dépasser la linéarité des textes et un relevé essentiellement statistique des faits collocationnels, par une prise en compte des contraintes distributionnelles, dans la tradition harrissienne, qui pèsent sur l'emploi des collocations dans les différents mouvements du texte. Notre étude montre que l'analyse distributionnelle tout particulièrement permet une caractérisation plus fine des comportements et des fonctions des collocations au sein de ces unités suprêmes de langage qui se croisent et néanmoins demeurent distinctes: discours, genre, texte.

3. Méthodologie

Nous avons développé une méthodologie d'analyse des textes adaptée à l'observation de la distribution des collocations à travers les discours, genres et textes, méthode que nous expliquons dans cette section. Elle conduit à une analyse discursive des collocations dont les résultats pourront trouver des applications dans le développement de ressources lexicales, telles que celles visées par le projet ARTES, que nous présentons en première sous-partie.

3.1. Création de ressources lexicales avec le projet ARTES

Le projet ARTES (Aide à la Rédaction de TExtes Scientifiques) a été conçu pour exploiter les divers aspects des études en terminologie et en langues de spécialité: d'une part l'aspect didactique visant la formation à la traduction spécialisée, à la rédaction en langues de spécialité et à la conception de ressources linguistiques, et, d'autre part, l'aspect linguistique visant à parfaire nos connaissances sur les lexiques et les discours spécialisés. Le projet permet par conséquent de faire le pont entre la recherche et l'enseignement en langues de spécialité, et plus spécifiquement en traduction spécialisée (Pecman et Kübler 2011).

Lancé en 2007 à l'université Paris Diderot par les chercheurs travaillant sur les langues de spécialité au sein de l'équipe CLILLAC-ARP, ce projet a abouti en 2011 à la création d'une base de données terminologiques et phraséologiques en ligne: la base ARTES. Cette base a été conçue de manière à permettre l'évolution de son architecture en fonction des avancées de nos recherches sur les langues de spécialité, et en fonction des besoins des divers utilisateurs visés par l'outil: traducteurs, chercheurs, experts, qu'ils soient confirmés ou apprenants.

De par la richesse des informations qu'elle cherche à fournir (cf. section 5 de cette étude), elle vise à proposer une ressource complémentaire aux banques de données terminologiques connues, telles que Termium, Grand Dictionnaire Terminologique ou Eurodicautom, désormais connu sous le nom IATE. La base ARTES est toutefois plus proche des initiatives pour la création de ressources linguistiques où l'enseignement, la recherche et la conception de ressources sont très étroitement liés, telles que le projet DiCoInfo, ou son cousin DiCoEnviro ou encore le projet WebTerm.

L'utilité indéniable des ressources consacrées aux langues de spécialité justifie pleinement l'effort visant à améliorer l'approche lexicographique de ce type de données lexicales par une analyse de leur fonctionnement discursif.

3.2. Constitution et interrogation du corpus

3.2.1. Sélection des textes

La constitution du corpus et la sélection des textes ont été menées en fonction des objectifs de cette étude qui sont:

- l'observation des phénomènes collocationnels au niveau d'un genre et d'un type de discours, en l'occurrence à l'intérieur des articles scientifiques;
- l'observation de la distribution des phénomènes collocationnels à l'intérieur d'un texte.

Les collocations spécifiques à un genre et à un type de discours peuvent être observées de manière efficace dans un corpus comportant des documents du même genre mais portant sur des domaines différents. Pour cette étude, nous avons choisi un corpus d'articles scientifiques issus de différentes disciplines des sciences de la terre, de l'environnement et de l'écologie. Notre hypothèse est qu'une collocation spécifique à un genre aura tendance à transcender les domaines et, par conséquent à se manifester à travers le corpus entier. Un tel corpus permettra aussi d'opposer les collocations qui se manifestent à travers un genre à celles qui sont caractéristiques d'un domaine ou d'un texte.

Une alternative pour typer les collocations spécifiques aux genres serait sans doute de constituer un corpus de textes relevant de genres différents mais portant sur un même domaine. Cela permettrait par exemple de faire ressortir les propriétés spécifiques aux différents genres (par exemple celles d'un article scientifique, par opposition à celles d'une communication orale, ou encore d'un article de presse à visée de vulgarisation). Cette méthode n'est pas explorée dans le présent article, étant donné que la base ARTES est orientée davantage vers les ressources destinées à la rédaction et la traduction d'articles scientifiques, mais elle pourra constituer une orientation de recherche dans le futur.

Considérant la finesse d'analyse à laquelle nous souhaitions parvenir, nous avons décidé d'utiliser deux corpus. Le premier est un corpus de très petite taille comportant 49.000 mots, soit dix articles scientifiques dont il a été possible de procéder à un découpage des textes (cf. 3.2.2.) et à une analyse détaillée, combinant analyse manuelle et automatique. Ce corpus est considéré comme le corpus principal. Il comporte des textes portant sur les sciences de la terre, l'environnement ou l'écologie qui sont majoritairement des publications récentes (huit articles de 2000-2011 et deux plus anciens de 1943 et 1986).

Pour certains types d'analyses, un corpus de taille plus importante a permis de vérifier les tendances observées dans le corpus principal. Nous avons ainsi utilisé également un corpus de 14.620.000 mots portant sur les différentes disciplines des sciences de la terre (e.g. volcanologie, tectonique des plaques, sismologie, glaciologie, etc.) pour confirmer certaines pistes d'interprétation. Nous nous référons à ce corpus dans la section analyse sous le terme de « corpus STEP, » tandis que le corpus principal est nommé « corpus IMRAD. »

Soulignons encore que les deux corpus recueillent des textes collectés grâce à des banques de données textuelles telles que ScienceDirect. Les textes ont été sélectionnés après consultation avec des chercheurs en STEP pour garantir leur représentativité de la langue de spécialité en question. Ces textes n'étant pas libres de droits, les corpus ne sont pas destinés à la diffusion. Nous exploitons donc uniquement les données textuelles extraites de ces articles.

3.2.2. Découpage des textes en mouvements rhétoriques

Afin de permettre une analyse plus fine du comportement des collocations, nous avons procédé au découpage des textes du corpus IMRAD en fonction des sections définies dans les articles. Nous nous sommes fondée sur la structure IMRAD des articles scientifiques pour élaborer un modèle de découpage qui permet de prendre en compte la structure classique des textes, mais également les différences parfois manifestes d'un article à l'autre. Les Figures 1 et 2 montrent le modèle suivi et l'organisation des textes en corpus et sous-corpus. Par rapport au modèle IMRAD classique

qui mentionne les parties clés des articles (Introduction, Méthodes, Résultats et Discussion) notre modèle détaille toutes les sous-parties potentiellement identifiables dans un article et prévoit une possibilité d'adaptation de ce modèle pour les articles qui ne se plient que partiellement à cette structure. Par exemple, un article qui n'indique pas de manière explicite les sous-parties méthodes, résultats et discussion (numérotées 4 à 6 dans notre modèle), mais qui comporte néanmoins plusieurs sous-parties entre l'introduction et la conclusion, est découpé et codé de façon à indiquer l'emplacement, le nombre et l'ordre de ces sous-parties dans l'ensemble de la structure du texte: 4-6_part1of5, 4-6_part2 of5, 4-6_part3of5, etc. (cf. l'illustration dans la Figure 1).

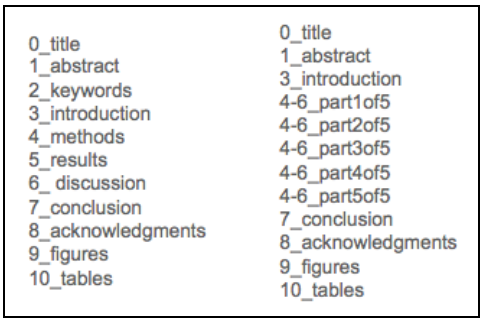

Figure 1: Modèle de découpage des textes en sections IMRAD

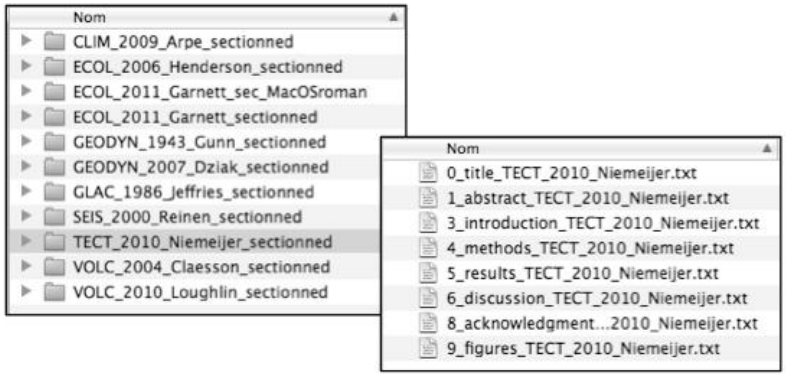

Figure 2: Constitution du corpus et des sous-corpus selon le modèle IMRAD

La présente étude constitue ainsi un essai de la prise en compte des mouvements des textes dans l'analyse de la distribution des collocations.

3.2.3. Interrogation du corpus

Pour interroger le corpus, nous avons tout d'abord procédé à une analyse manuelle des dix textes de notre corpus principal (cf. 3.2.1) afin de repérer les cas potentiels de collocations significatives. Dans un deuxième temps, nous avons tiré parti de la rigueur des outils d'analyse automatique pour vérifier si les phénomènes observés manifestent une distribution répartie au niveau du domaine, du genre, ou plutôt à l'intérieur d'un texte.

Nous avons choisi d'interroger le corpus à l'aide du logiciel Textométrie (TXM), qui intègre les dernières technologies en matière de TAL et qui permet de formuler des requêtes en Corpus Query Language (CQL). L'un des avantages de ce langage est de pouvoir extraire des listes de concordances pour des schémas lexico-grammaticaux complexes, tels que:

```
[word="provide.*"][]{0,10}[word="information|explanation.*|indication.*|insight.*"]
```

et par conséquent de lister des contextes comportant des collocations aussi diverses que:

> *to provide information to/that, to provide an alternative explanation of, to provide some insights into, to provide some indication of,* etc.

La méthode, combinant une analyse manuelle et automatique, permet ainsi de formuler les hypothèses sous forme de requête et de vérifier la validité de ces mêmes hypothèses dans l'ensemble du corpus. Parmi les nombreuses structures collocationnelles observées, nous présentons ici quelques exemples.

4. Analyse distributionnelle et lexicographique des phénomènes collocationnels

Les résultats de nos observations du comportement des collocations dans les articles scientifiques sont illustrés par des exemples représentatifs de ces comportements. L'analyse consiste en un essai de caractérisation des collocations quant à leur comportement discursif. Dans la dernière section de cet article, nous montrons les applications possibles de ce type de recherche pour le traitement lexicographique des collocations tel qu'il a été mis en place dans la base ARTES.

Deux grands types de comportement ont pu être relevés: les collocations dont la distribution marque un fonctionnement au niveau du genre, et celles dont la distribution présente davantage un ancrage à l'intérieur soit d'un texte soit d'un domaine. Les premières sont par ailleurs liées à un type de discours – en l'occurrence, dans le cas de cette étude, au discours scientifique – et les dernières sont liées à un type de connaissances – dans le cas de cette étude il s'agit des connaissances liées aux domaines représentés par le corpus.

4.1. Analyse distributionnelle des collocations spécifiques au genre et au type de discours

Il s'agit de repérer les collocations qui pourraient être conditionnées par le genre textuel ou type de discours. Chaque communauté de discours a ses genres textuels privilégiés. Par exemple, l'article scientifique, la communication à un colloque, la monographie, la contribution à un

ouvrage, le rapport sur un projet de recherche, la thèse, l'HDR, l'abstract, etc. sont autant d'exemples de genres caractéristiques des diverses communautés scientifiques. Chaque genre textuel propose un cadre ou schéma de communication et des marqueurs linguistiques qui permettent de l'identifier. Les genres vont souvent de pair avec les types de discours qui eux désignent l'ensemble des productions langagières d'une communauté linguistique ciblée, tels que les discours scientifique, administratif, politique, médiatique, juridique, etc. Un type de discours est donc lié à une langue de spécialité – dans la lignée de l'opposition entre langue et parole ou langue et discours – tandis qu'un genre est la forme sous laquelle se réalise un discours.

Les collocations présentant une cooccurrence des items lexicaux à l'intérieur d'un genre et d'un type de discours peuvent être caractérisées comme « génériques » du point de vue de leur contenu sémantique, dans la mesure où elles ne servent pas à désigner une information ou connaissance d'un domaine spécifique: par ex. *to provide explanation, to provide insight, we further thank anonymous reviewers for, to thank sb for their constructive comments*, etc.

La Figure 3 montre à titre d'exemple une requête formulée à partir d'un cas de collocations lexicales (par opposition aux collocations grammaticales dont un exemple sera discuté ci-après), i.e. dont les items cooccurrentiels sont tous les deux de nature lexicale: *to provide information/explanation/insight/indication.*

Leur présence et leur distribution à travers différentes disciplines suggère un emploi spécifique au genre (article) et type de discours (scientifique).

text_id	Contexte gauche	Pivot	Contexte droit
1_abstract_TECT_2010_Niemeijer	our observations may	provide a common explanation	for weakness of mature faults
5_results_VOLC_2004_Claesson	sandstone aquifers might	provide an alternative explanation	for Na / Ca ratio
1_abstract_SEIS_2000_Reinen	deformed serpentinite gouges	provide information	that may be used to
7_conclusion_SEIS_2000_Reinen	be used to	provide insight	into the seismic behavior of
5_results_VOLC_2004_Claesson	1969),	provide some insights	into possible source – switching
4–6_part3of6_CLIM_2009_Arpe	Fig. 2a	provides some indication	of the impacts of global
7_conclusion_CLIM_2009_Arpe	possible and to	provide the end – users with the best available information	to guide their mitigation and

Figure 3: Cooccurrence de *provide* avec *information, explanation, indication, insight* dans le corpus IMRAD

L'interrogation du corpus des STEP, de taille plus importante, a permis dans tous les cas de corroborer nos interprétations, par le nombre important d'occurrences correspondant à des schémas collocationnels que nous avions identifiés dans notre corpus IMRAD. La Figure 4 montre, à titre d'exemple, un échantillon des 870 occurrences du même schéma dans le corpus STEP.

text_id	Contexte gauche	Pivot	Contexte droit
LELOUTRE_GL_Crosetto_2005	the flexibility to	provide (qualitative) information	on deformations,
ESSAKOMBE_SI_2005_MALONE	and research to	provide accurate information	to reporters.
MARTIKAINEN_GE_2003_Dussauge	heterogeneous conditions to	provide accurate information	on the fracture
LELOUTRE_VL_Byrne_2009	derived slope maps	provide a clear insight	into terrace plan
MENDES_VL_2008_ross	driven material,	provide a compelling explanation	for the origin
PANTOLI_GE_2000_FORT	this does not	provide a complete explanation	because it is
FUJIMOTO_PT_2006_Bendick	only intended to	provide a conceptual explanation	for the temporal
PINTEA_OC_2002_Bachu	, these distributions	provide a correct indication	of the strength
NGUYEN_GL_2009_Roberts	, our results	provide a crude indication	of ice surface
RICHARD_SI_2002_HAZZARD	the model can	provide added information	not available from
MORELLE_GL_2006_Anschutz	previous studies and	provide additional detailed information	about smallscale spatial
LAZAR_ES_2004_Müntener	element chemistry should	provide additional indications	of whether considerable
BISIAUX_VL_2007_Geyer	(1984)	provide additional information	regarding caldera areas
BLONDEL_VL_2008_Geyer	(1984)	provide additional information	regarding caldera areas
CHEVALIER_PT_2006_Wolfenden	and geomorphological data	provide additional information	on the zone
CLAVEL_GE_2009_Bor	the present array	provide additional information	to constrain the
CORNU_GE_2000_Wasowski	the 1984 earthquake	provide additional information	relevant to hazard
LELOUTRE_GL_VL_Geyer_2008	(1984)	provide additional information	regarding caldera areas
MILSTEIN_SI_2005_PELEGRI	objective is to	provide additional information	that may help
NGUYEN_GL_2010_Polyak	This history will	provide additional information	on the stability
RIBES_PD_2003_Nour	and does not	provide additional information	of practical use
RICHARD_SI_2004_LATORRE	data may also	provide additional information	on sharp interfaces
THORAVAL_PT_2004_Wolfenden	and geomorphological data	provide additional information	on the zone
NGUYEN_GL_2010_Polyak	grains, can	provide additional insight	([Helland
LAGEDAMOND_GE_2008_Mouslopoulou	, the models	provide a first – order indication	of what earthquake
GUILLAUMEAU_GL_2010_Nettles	of glacier termini	provide a general explanation	for the variability
PANTOLI_GE_2010_NETTLES	of glacier termini	provide a general explanation	for the vari
PANNETIER_GE_2002_Dai	are able to	provide all the information	required for landslide

Figure 4: Cooccurrence de *provide* avec *information, explanation, indication, insight* dans le corpus STEP

Les collocations relevées dans le corpus STEP permettent de faire l'hypothèse d'un schéma plus complexe et plus détaillé que celui fondé sur le corpus IMRAD:

> *to provide (additional/accurate/qualitative/detailed/all the) information on/to; to provide (general/complete/conceptual/compelling) explanation for; to provide (correct/crude) indication of; to provide (additional/clear) insight into*

Le corpus IMRAD offre l'avantage de pouvoir prendre en compte les caractéristiques distributionnelles des collocations. Le comportement de ce type de collocations selon les différentes sections des articles, présenté dans la Figure 3, suggère un emploi restreint à des parties qui servent davantage à présenter les aboutissements d'une étude ou recherche, tels qu'abstract, résultats et conclusion. Cette distribution concorde avec le contenu sémantique de ces collocations qui servent à formuler une interprétation des résultats et de leur utilité pour la compréhension des phénomènes observés (cf. par exemple le Tableau 2 pour l'étiquetage de ce type de collocations dans la base ARTES selon leur fonction discursive).

La Figure 5 illustre une requête formulée à partir d'un cas de collocation grammaticale, i.e. dont l'un des items cooccurrentiels est de nature lexicale et les autres de nature grammaticales (pronom et préposition en l'occurrence): *we/I thank … for* :

text_id	Contexte gauche	Pivot	Contexte droit
8_acknowledgements_CLIM_2009_Arpe	, 1972). We	thank	PAGES and IGCP490 for supporting the attendance to the GCAGS –
8_acknowledgements_CLIM_2009_Arpe	this manuscript. We further	thank	the anonymous reviewers for their constructive and helpful comments. Ref
8_acknowledgments_TECT_2010_Niemeijer	Calgary. Ô³øWe wish to	thank	Bob Holdsworth and an anonymous reviewer for their helpful suggestions
8_acknowledgments_TECT_2010_Niemeijer	We would also like to	thank	Steve Swavely for his technical support and Brett Carpenter and Cristiano
8_acknowledgments_VOLC_2004_Claesson	is gratefully acknowledged. We	thank	H. Siegmund and K. Hajnal (Stock – holm
8_acknowledgments_VOLC_2010_Lauglin	23] Acknowledgments. We	thank	staff and volunteers at the MVO for technical assistance and support
8_acknowledgmet_SEIS_2000_Reinen	(NERC). I	thank	E. Grosfils, G. Gleason, P. Santanach
8_acknowlidgments_GEODYN_2007_Dziak	. Resing. We also	thank	S. Carbotte and an anonymous reviewer for very helpful comments
8_acknowlidgments_GEODYN_2007_Dziak	J. Klay. We	thank	the National Oceanic and Atmospheric Administration (NOAA) Vents Pro

Figure 5: Collocations de *we/I thank … for* dans le corpus IMRAD

L'hypothèse d'un schéma sous-jacent pourrait être formulée de la manière suivante:

> *I/we (wish to/would like to) (also/further) thank sb/anonymous/reviewers/staff/volunteers for (their/his/her) (very) (helpful/constructive) comments/suggestions/technical assistance/support*

D'un côté, la distribution à travers différents textes des collocations présentée dans la Figure 5 suggère qu'il s'agit bien d'un schéma qui transcende les disciplines – en d'autres termes qu'il s'agit bien des collocations spécifiques à un genre et un type de discours. De l'autre côté, la distribution très peu étendue permet d'affirmer que l'emploi de ce type de schéma est très codé et réservé à la section des remerciements.

L'analyse du cas des collocations spécifiques au genre montre qu'elles peuvent avoir une distribution restreinte à travers les différentes sections des articles et que, de ce fait, elles constituent un élément connu et attendu de la part du lecteur, qui fait partie de la même communauté scientifique de locuteurs que l'auteur. Comme le fait remarquer Bordet (2011), l'utilisation de ces mêmes formules stéréotypées participe à la construction d'une sorte de zone de proximité avec le lecteur, qui doit reconnaître dans l'article scientifique son propre univers de discours. Par conséquent, elles jouent un rôle cohésif très fort au niveau du genre.

4.2. Analyse distributionnelle des collocations à l'intérieur d'un texte

L'analyse distributionnelle des collocations à l'intérieur d'un texte permet d'illustrer le comportement et le rôle des collocations au niveau de la construction du texte. Nous prenons comme exemple les combinaisons récurrentes entre l'item lexical *weak* – pouvant être caractérisé comme générique du point de vue de son contenu sémantique – avec des termes qui renvoient à un des concepts spécifiques du domaine de la minéralogie et de la tectonique des plaques: *talk, clay, phyllosilicate, foliation, fault,* etc. Afin de prendre en compte toutes les variations potentielles, nous avons lancé la recherche à partir du morphème *weak*; ce qui a permis de relever tous les cas de dérivation.

Les Figures 6 et 7 montrent une très forte concentration du morphème *weak* (44 occurrences au total) qui, dans un même texte, présente deux

caractéristiques. La première est sa forte tendance à la variation dérivationnelle: *weak, weaker, to weaken, weakened, weakening, weakness*. La deuxième est sa régularité dans les combinaisons avec des items spécifiques du domaine: *mineral, clay, talc, layer, interlayer, foliation, fault*, etc.

text_id	Contexte gauche	Pivot	Contexte droit
0_title_TECT_2010_Niemeijer	Ôªøfabric induced	weakness	of tectonic faults ÔªøMature fault
1_abstract_TECT_2010_Niemeijer	fault zones appear to be	weaker	than predicted by both theory
1_abstract_TECT_2010_Niemeijer	explanation involves the presence of	weak	minerals, such as talc
1_abstract_TECT_2010_Niemeijer	: what proportion of a	weak	mineral is needed to satisfy
1_abstract_TECT_2010_Niemeijer	mineral is needed to satisfy	weak	fault models ? Existing studies
1_abstract_TECT_2010_Niemeijer	> 30 % of the	weak	phase is necessary to weaken
1_abstract_TECT_2010_Niemeijer	weak phase is necessary to	weaken	faults , ê £ a
1_abstract_TECT_2010_Niemeijer	. Here we demonstrate that	weakening	of fault gouges can be
1_abstract_TECT_2010_Niemeijer	provide a common explanation for	weakness	of mature faults. The
3_introduction_TECT_2010_Niemeijer	mature faults. The apparent	weakness	of mature faults with large
3_introduction_TECT_2010_Niemeijer	explained by extreme fault zone	weakness	[Collettini and Sibson,
3_introduction_TECT_2010_Niemeijer	]. Possible explanations for	weak	faults include low effective stress
3_introduction_TECT_2010_Niemeijer	], the presence of	weak	phyllosilicates or clays [Imber
3_introduction_TECT_2010_Niemeijer	, 2009a], dynamic	weakening	[Melosh, 1996 ;
3_introduction_TECT_2010_Niemeijer	, the origin of fault	weakness	remains poorly understood and a
3_introduction_TECT_2010_Niemeijer	led to speculations that the	weakness	of these faults can be
3_introduction_TECT_2010_Niemeijer	talc is needed to satisfy	weak	fault models ? Recent experiments
3_introduction_TECT_2010_Niemeijer	layer, which would significantly	weaken	the composite gouge as long
3_introduction_TECT_2010_Niemeijer	that fabric can induce significant	weakness	even under cataclastic conditions.
3_introduction_TECT_2010_Niemeijer	, 2006] show that	weakening	can occur with as little
3_introduction_TECT_2010_Niemeijer	10 wt % of the	weak	mineral, provided that the
3_introduction_TECT_2010_Niemeijer	ê £ solution and the	weak	mineral forms a through ,

Figure 6: Occurrences du morphème *weak* dans le corpus IMRAD

text_id	Contexte gauche	Pivot	Contexte droit
3_introduction_TECT_2010_Niemeijer	stress, provided that a	weak	and continuous foliation is present
3_introduction_TECT_2010_Niemeijer	the role of fabric in	weakening	due to the plasticity and
3_introduction_TECT_2010_Niemeijer	of material required to induce	weakening	under brittle, cataclastic conditions
5_results_TECT_2010_Niemeijer	stress followed by a gradual	weakening	. In contrast, the
5_results_TECT_2010_Niemeijer	with a talc interlayer were	weaker	than the homogeneous mixtures (
5_results_TECT_2010_Niemeijer	, followed by gradual strain	weakening	(Figure 1).
6_discussion_TECT_2010_Niemeijer	and indicates the onset of	weakening	relative to pure quartz (
6_discussion_TECT_2010_Niemeijer	a fully , ê £	weakened	layer, with the frictional
6_discussion_TECT_2010_Niemeijer	talc. One expects that	weakening	may vary with layer normal
6_discussion_TECT_2010_Niemeijer	work. The degree of	weakening	depends on the thickness of
6_discussion_TECT_2010_Niemeijer	on the thickness of the	weak	layer. In nature,
6_discussion_TECT_2010_Niemeijer	, talc and most other	weak	minerals (clays and phyllosilicates
6_discussion_TECT_2010_Niemeijer	mature fault zones without inherited	weak	minerals, should be frictionally
6_discussion_TECT_2010_Niemeijer	reac – tions produce the	weak	phases in , ê £
6_discussion_TECT_2010_Niemeijer	minerals, ultimately forming a	weak	through , ê £ going
6_discussion_TECT_2010_Niemeijer	. Moreover, observations of	weak	minerals in fault zones coupled
6_discussion_TECT_2010_Niemeijer	that even trace quantities of	weak	phyllosilicates can induce fault weakness
6_discussion_TECT_2010_Niemeijer	weak phyllosilicates can induce fault	weakness	if the minerals form a
6_discussion_TECT_2010_Niemeijer	develop the requisite network of	weak	minerals that may provide reduced
6_discussion_TECT_2010_Niemeijer	only trace concentrations of the	weakening	phase. These ingredients are
9_figures_TECT_2010_Niemeijer	thickness for the onset of	weakening	and for , Äúmaximum ,
9_figures_TECT_2010_Niemeijer	for , Äúmaximum , Äù	weakening	are indicated. These thresholds

Figure 7: Suite des occurrences du morphème *weak* dans le corpus IMRAD

Les résultats d'interrogation du corpus permettent de proposer plusieurs éléments d'interprétation pour ce cas. Tout d'abord, le schéma collocationnel sous-jacent de *weak* repose sur le phénomène de métonymie par l'association de cet item avec des termes désignant une substance – *mineral*, *clay*, *talc* ou *phyllosilicate* – ou une formation contenant cette substance – *layer, interlayer, phase, foliation, gouge* – et finalement par la localisation de cette formation sur un édifice géologique, en l'occurrence *fault*; ce qui explique qu'on puisse trouver des combinaisons du type: *weak mineral*, *weak layer* ou *weak fault*. Ce schéma pourrait être formulé en CQL de la manière suivante:

[word="weak"] [sem="substance" OR word="mineral|clay|talk|phyllosilicate"] |[sem="formation" OR word="layer|interlayer|phase|foliation|gouge"]| [sem="location" OR word="fault"]

Ensuite, le schéma collocationnel repose sur une distribution entre les formes dérivées de base (*weak, weaker, weaken, weakened*) et les formes résultant de la nominalisation (*weakening* et *weakness*) dans la mesure où les premières marquent une tendance à être suivies par un nom, tandis que les dernières se combinent davantage avec un certain type de verbes (*to induce…* ou *…can occur, may vary, depends on*) et admettent une caractérisation de type adjectival (*significant, apparent, induced, gradual, dynamic*…). Le Tableau 1 permet de résumer ces tendances.

<table>
<tr><td></td><td>weak</td><td>mineral, talc, clay, phyllosilicate
layer, phase, foliation, gouge
fault</td></tr>
<tr><td></td><td>weaker</td><td>talc interlayer
fault zones</td></tr>
<tr><td></td><td>weaken</td><td>fault, gouge</td></tr>
<tr><td></td><td>weakened</td><td>layer</td></tr>
<tr><td>to induce
gradual, dynamic</td><td rowspan="2">weakening</td><td></td></tr>
<tr><td></td><td>can occur
may vary
depends on</td></tr>
<tr><td>fault (zone)
to induce
significant, apparent, induced</td><td rowspan="2">weakness</td><td></td></tr>
<tr><td></td><td>of (these/mature) faults</td></tr>
</table>

Tableau 1: Schéma collocationnel du *weak* et de ses dérivés

L'élément complémentaire d'analyse porte sur la distribution des collocations du morphème *weak* qui se révèle être très forte – titre, abstract, introduction, résultats, discussion, figures – (la seule section non représentée étant la méthode). On peut donc penser que les collocations, dans leur variation et leur unicité, assurent la cohésion et la progression au niveau du texte. Les structures collocationnelles variées permettent de déplacer la focalisation d'un mouvement de texte à l'autre.

Enfin, un dernier élément d'analyse porte sur le recours systématique à la variation à travers les dérivés de *weak*. Ce transfert d'un item lexical générique à travers les catégories grammaticales repose sur le mécanisme de « métaphore grammaticale » (cf. Halliday 1998) qui, dans le discours scientifique, permet de problématiser une notion *a priori* générique et par là de la glisser dans un domaine spécialisé.

4.3. Analyse des collocations spécifiques à un domaine de connaissances

Il est tout à fait possible de relever des combinaisons lexicales qui n'ont pas forcément un fonctionnement saillant à l'intérieur d'un texte, ni une distribution qui pousserait à leur attribuer un rôle dans la reconnaissance d'un genre textuel, mais qui portent sur des connaissances spécifiques à un ou plusieurs domaines. Ce sont d'ailleurs les collocations le plus souvent observées dans les études en langues de spécialité, ces études s'appuyant sur des corpus de domaines.

Dans notre corpus principal, nous pouvons relever le cas du verbe *accommodate* qui se combine avec d'autres termes du domaine. La Figure 8 montre les concordances de ce terme verbal dans le corpus IMRAD. Employé d'habitude dans la langue générale, ce verbe pourrait passer inaperçu. Or, il s'agit bien d'un terme du domaine qui, dans la structure *to accommodate a slip*, renvoie à un sens particulier en tectonique des plaques. On est ici en présence d'un glissement du sens premier très général (« adapter, arranger qqch ») vers un sens plus spécialisé (« action de rétablissement de l'équilibre tectonique suite à un mouvement principal par les mouvements secondaires qui cherchent à repositionner les plaques le plus favorablement possible »). Le schéma de combinaison *to accommodate a liquid into sth* relève du domaine de la géodynamique et montre une autre acception de ce terme dans un domaine connexe.

text_id	Contexte gauche ▲	Pivot	Contexte droit
6_discussion_TECT_2010_Niemeijer	to 10s of meters wide and have	accommodated	kilometers of slip during their lifetime [
4-6_part3of4_SEIS_2000_Reinen	Thus in regime I, deformation is	accommodated	by cataclastic flow of the gouge,
6_discussion_GEODYN_2007_Dziak	. This new, hotter liquid is	accommodated	into the finite volume of the preexisting
6_discussion_GEODYN_2007_Dziak	. Nevertheless, once new liquid is	accommodated	into the finite volume of the preexisting
3_introduction_SEIS_2000_Reinen	laboratory studies have demonstrated that it may	accommodate	both aseismic and seismic slip under shallow

Figure 8: Collocation de *accommodate* et de ses dérivés dans le corpus IMRAD

Le corpus STEP confirme cette tendance avec 1096 occurrences de *accommodate* et ses dérivés dont un échantillon, présenté à la Figure 9, permet d'avancer le schéma suivant: *to accommodate a slip/deformation/extension*, ou la version nominalisée: *the accommodation of a slip/deformation/extension*.

text_id	Contexte gauche	Pivot	Contexte droit
BERTEA_SI_2005_Rice	is easier to	accommodate deformation	by slip between
BERTEA_SI_2005_Rice	appear to have	accommodated most of the slip	, and which
BISIAUX_GE_2006_Acocella	dipping normal faults	accommodate the gravitational deformation	. Multiple concentric
BISIAUX_VL_2002_Cole	instabilities cannot be	accommodated by further extension	). Alternatively
BISIAUX_VL_2005_Hill	1989 was entirely	accommodated by minor deformation	of the surrounding
BLONDEL_VL_2007_Bonforte	underlying basement,	accommodating the deformation	along the southern
BLONDEL_VL_2007_Bonforte	volcanic cover,	accommodates the plastic deformation	of the sedimentary
CHEVALIER_OC_2005_Kriner	are required to	accommodate the increased extension	. The additional
CHEVALIER_PT_2002_Eagles	– bounding faults	accommodating extension	strike north.
CHEVALIER_PT_2008_Kinnaird	Fault Zone)	accommodates deformation	internal to the
CHEVALIER_PT_2008_Kinnaird	earthquakes, which	accommodate reverse – slip	as oceanic crust
CHEVALIER_PT_2008_Kinnaird	earthquakes, which	accommodate normal and strike – slip	faulting in the
CHEVALIER_PT_2008_Kinnaird	which exist to	accommodate horizontal slip	between sections of
CHEVALIER_PT_2008_Kinnaird	– slip faulting	accommodating horizontal slip	between sections of
CHEVALIER_PT_2008_Kinnaird	of faulting was	accommodated by differential extension	in the overriding
CLAIRE_GE_2007_Frankel	Valley fault zones	accommodate the majority of slip	in the ECSZ
CLAIRE_GE_2007_Frankel	Valley fault zone	accommodates the majority of slip	in the region
CLAIRE_PT_2004_Chiarabba	and Eurasia is	accommodated by a complex deformation	of the Alps
CLAUDON_GE_2005_BURCHFIELpart6	the deformation is	accommodated by slip	on bedding or
CLAUDON_SI_2009_ZIELKE	favorably oriented to	accommodate extension	, and they
CLAUDON_SI_2009_ZIELKE	favorably oriented to	accommodate extension	, and older
CORNU_GE_2002_Trenkamp	diffuse plate boundary	accommodating deformation	over a broad

Figure 9: Collocation de *accommodate* et de ses dérivés dans le corpus STEP

Toutefois le faible nombre d'occurrences de ce schéma dans le corpus IMRAD ne nous permet pas d'offrir une caractérisation pertinente de leur comportement du point de vue distributionnel, autre que d'observer une tendance à figurer dans les sections liminaires des articles, notamment dans la partie discussion et, dans une moindre mesure, dans la partie introduction.

5. Applications au traitement lexicographique des collocations dans la base ARTES

Afin de fournir des ressources utiles pour la traduction et la rédaction en langue de spécialité, la base ARTES répertorie non seulement des termes mais également des expressions plus ou moins stéréotypées autour de ces termes ainsi que des formulations d'argumentation fréquemment utilisées dans les textes spécialisés. Elle permet de stocker des informations lexicales sans restriction quant aux domaines ou aux langues. Chaque entrée de la base comporte un ensemble de rubriques permettant de renseigner les utilisateurs sur l'emploi des unités linguistiques, telles que les définitions, les synonymes, les contextes, les collocations, les équivalents, etc. En complément de la terminologie, la base accorde ainsi une place importante au contexte et à la phraséologie, qu'il s'agisse de la phraséologie spécifique aux domaines ou de la phraséologie transversale, plus caractéristique des genres et des types de discours. Par ailleurs, la base ARTES est munie de deux applications: l'une pour l'édition et la gestion de la base, et l'autre pour la consultation des données en libre accès.

La consignation des collocations dans une base de données, dont le but est d'offrir des ressources pour la traduction et la rédaction en langue de spécialité, pose des problèmes lexicographiques majeurs. Dans nos travaux précédents (cf. Pecman 2008), nous avons déjà exploré certains aspects du traitement lexicographique des collocations, notamment du point de vue de l'accès offert aux ressources ou des problèmes posés par

la lemmatisation des collocations. Si les problèmes lexicographiques de présentation des ressources collocationnelles demeurent d'actualité, les projets tels que ARTES permettent d'explorer les pistes pour les résoudre.

La base ARTES reflète ainsi les apports de l'approche discursive des collocations dans les choix lexicographiques qui ont été faits pour intégrer les collocations à des ressources terminologiques. A notre connaissance, le projet ARTES est unique dans le sens où il accorde une importance égale à la terminologie et à la phraséologie des langues de spécialité, et où il cherche à caractériser cette dernière tant du point de vue des domaines, que des genres et des types de discours.

Dans l'état actuel de nos recherches, menées dans le cadre de ce projet, les collocations spécifiques aux domaines sont consignées dans les fiches terminologiques. Elles sont donc liées aux termes, et par voie de conséquence elles sont catégorisées en fonction des domaines. Par exemple, la collocation to *accommodate a slip* devra être enregistrée dans la fiche terminologique de chacun des termes la composant: *accommodate* et *slip*.

Les collocations spécifiques aux genres sont consignées indépendamment des domaines et des termes. Dans la terminologie ARTES, elles sont appelées « collocations génériques » et elles sont associées à un type de discours dominant – scientifique, technique, administratif, socio-économique, politique ou médiatique, etc. – et à l'une des « fonctions discursives » (cf. la partie 2.2 pour une définition de ce concept) dégagées dans une analyse préalable des textes scientifiques (Pecman 2007) et dont certaines sont offertes, à titre d'exemple, dans le Tableau 2. Le tableau montre également que les collocations génériques peuvent prendre des formes très diverses.

Collocations génériques	**Fonction discursive associée**
to provide (some) insights into	Décrire, interpréter et analyser les données ou phénomènes observés
to thank sb for very helpful comments	Exprimer des remerciements
the most complete account of (sth) is found in	Faire un renvoi à une personne ou une œuvre connue
these findings are first to describe	Parler des résultats et des découvertes de manière positive
my concern here is with	Annoncer le sujet de la section courante
there is little doubt (that)	Exprimer une incertitude ou une atténuation
as previously mentioned	Faire un renvoi à une partie dans le discours en cours
in much the same manner	Exprimer une compatibilité, une

	corrélation, une analogie ou une similitude
by contrast	Souligner une différence, une dissemblance ou une opposition
in this paper we survey the state of art in	Evoquer le sujet de son étude

Tableau 2: Exemple de collocations génériques associées à diverses fonctions discursives

Dans l'état actuel des recherches, quelques 80 fonctions discursives ont été répertoriées. Une telle classification offre une analyse plus fine et plus graduée que celles qui reposent sur la structure IMRAD des articles scientifiques. Les fonctions discursives constituent le point d'accès aux collocations génériques depuis l'application de consultation de la base ARTES (cf. Figure 10, encadré à gauche).

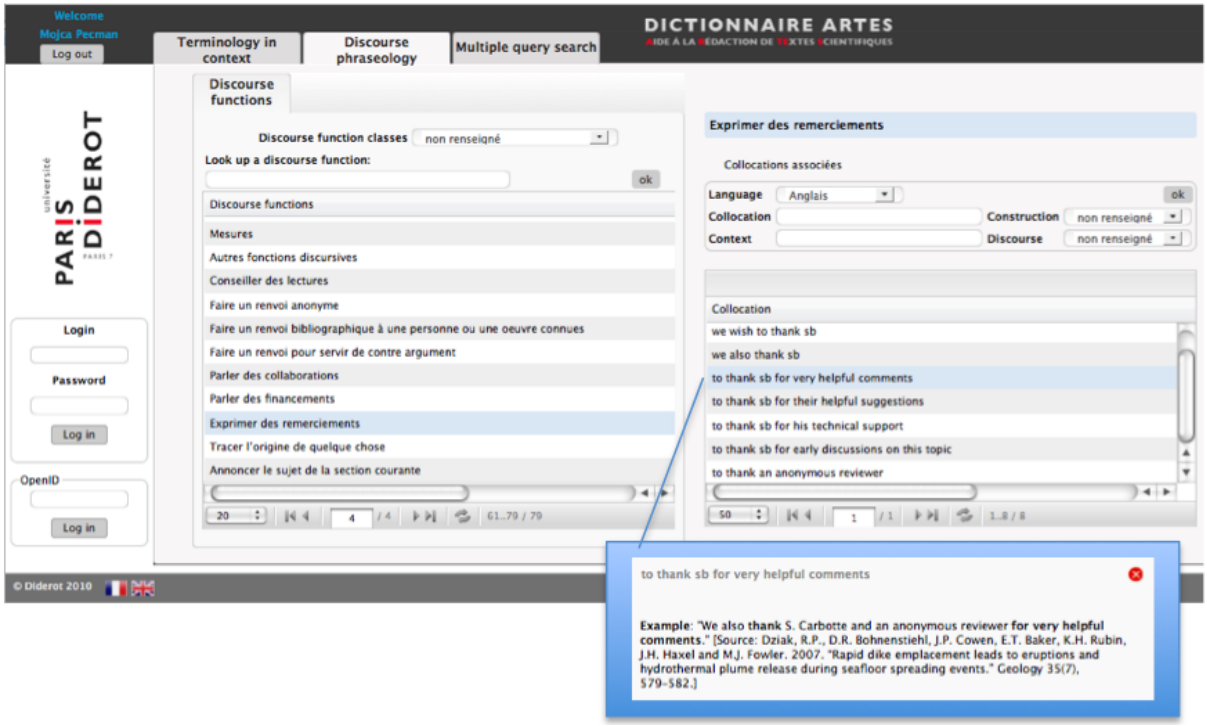

Figure 10 : Interface du dictionnaire de consultation de la base ARTES montrant un accès aux collocations génériques par fonctions discursives

A ce stade du projet, la méthode de consignation des collocations consiste à offrir pour chaque collocation un exemple illustrant l'usage de la collocation en contexte. De ce fait, les collocations sont enregistrées en tant qu'actualisations observées dans les textes en suivant la méthode de découpage en unités lexico-grammaticales minimales, à partir desquelles il est possible de reconstruire différents schémas (cf. Figure 10, encadré à droite).

Une analyse et réorganisation des données pour reconstruire des schémas lexico-grammaticaux sous-jacents restent donc néanmoins une

perspective de ce type d'approche des ressources. Les collocations servant à exprimer un remerciement pourraient ainsi être présentées sous la forme suivante:

> *I/we (wish/would like to) (also/further) thank sb/anonymous reviewers/staff/volunteers for (their/his/her)? (very)? (helpful/constructive)? comments/suggestions/technical assistance/support*

Nous pensons que la prise en compte des dimensions du discours, du genre et du texte dans les études sur des collocations, pourrait encore nous aider à améliorer leur traitement lexicographique dans la base ARTES.

6. Conclusion

Loin d'être un simple fait statistique, les collocations, en tant qu'associations privilégiées de deux ou plusieurs unités linguistiques, assurent des fonctions discursives qu'il est possible de caractériser à travers une analyse distributionnelle des collocations. Cette étude met en évidence le rôle spécifique des collocations au niveau du genre, du type de discours, d'un domaine de connaissances et d'un texte clos. D'ordinaire, les collocations dans les langues de spécialité sont considérées presque exclusivement à l'intérieur d'un domaine, et interprétées comme éléments caractéristiques de ce même domaine de connaissances. Notre étude montre que changer de prisme d'observation, pour se focaliser sur le genre ou sur un texte clos, ouvre des voies pour une nouvelle interprétation des phénomènes collocationnels. Certains comportements, au niveau du genre notamment, montrent la capacité des collocations à assurer, à travers un texte, l'identification à une communauté de discours et un mode de communication. Ainsi, la réutilisation de formules stéréotypées dans les articles scientifiques (e.g. *the outcome/findings/results of this study suggest*; *in regard to, in regards to, in this regard*) fonctionne comme un élément cohésif du genre. A l'intérieur d'un texte, les collocations permettent d'assurer à la fois sa cohésion et sa progression, par les reprises et les variations qui ponctuent les différents mouvements textuels (e.g. *weak layer > weak fault > weaker talc interlayer > to induce weakness of faults*). L'analyse discursive des collocations offre ainsi une approche très fine des phénomènes cooccurrentiels. Elle permet d'améliorer nos connaissances sur ce type d'unité linguistique et nos méthodes de création de ressources collocationnelles. La base de donnée ARTES, qui intègre la terminologie et la phraséologie spécialisées, permet d'expérimenter la création de ressources phraséologiques où l'on opère la distinction entre les collocations spécifiques aux domaines (e.g. *to accomodate a slip, to weaken a fault*) et celles spécifiques à un type de genre et de discours (e.g. *to provide insigts into sth, we wish to further thank sb, to thank anonymous reviewers for their useful comments*). Enfin, cette approche

lexico-discursive des collocations permet de construire des ressources plus adaptées à la traduction et la rédaction en langue de spécialité.

Remerciements

Je saisis ici l'occasion de rendre hommage et d'exprimer ma reconnaissance envers le Professeur Henri Zinglé qui m'a fait découvrir la phraséologie et les approches de la linguistique appliquée qui, depuis, nourrissent mes recherches en langues de spécialités. Je tiens à remercier également Geneviève Bordet et deux relecteurs anonymes de la revue JoSTrans pour leurs remarques constructives qui ont permis d'améliorer sensiblement cet article.

Références bibliographiques

- **Benson, Morton** (1989). "The Structure of the Collocational Dictionary." *International Journal of Lexicography* 2(1), 1-14.
- **Benson, Morton, Evelyn Benson and Robert Ilson** (1997). *The BBI Dictionary of English Word Combinations*. 2nd ed. Amsterdam/Philadelphia: John Benjamins.
- **Biber, Douglas, Ulla Connor et Thomas Albin Upton** (2007). *Discourse on the move*. Amsterdam/Philadelphia: John Benjamins.
- **Bordet, Geneviève** (2011). *Etude contrastive de résumés de thèse dans une perspective d'analyse de genre.* Thèse de doctorat. Université Paris 7-Paris Diderot.
- **Coxhead, Averil** (2002). "The Academic Word List: A Corpus-based Word List for Academic Purposes." *Teaching and Language Corpora (TALC) 2000 Conference Proceedings.* Atlanta: Rodopi, 73–89.
- **Coxhead Averil et David Hirsh** (2007). "A pilot science-specific word list." *Revue française de linguistique appliquée* 12(2), 65-78.
- **de Beaugrande, Robert** (1980). *Text, Discourse and Process*. Norwood: Ablex.
- **Dubreuil, Estelle** (2008). "Collocations: Définitions et problématiques." *Texto!* 13(1), 1-39.
- **Firth, John Rupert** (1968). *Selected Papers of J.R. Firth 1952-59*. Frank Robert Palmer (ed.) (1968) London/Harlow: Longman.
- **Fontenelle, Thierry** (1994). "Towards the construction of a collocational database for translation students." *Meta : journal des traducteurs* 39(1), 47-56.
- **Gledhill, Christopher** (2000). *Collocations in Science Writing.* Language in performance, 22. Tuebingen: Gunter Narr Verlag.
- — (2011). "The 'lexicogrammar' approach to analysing phraseology and collocation in ESP texts." *Asp* 59, 5-23.
- **Granger, Sylviane** (1998). "Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae." Anthony Paul Cowie (ed.) (1998) *Phraseology: Theory,*

Analysis, and Applications. Oxford: OUP, 145–160.

- **Granger, Sylviane and Magali Paquot** (2008). "Disentangling the phraseological web." Sylviane Granger and Fanny Meunier (eds) (2008). *Phraseology: An Interdisciplinary Perspective*. Amsterdam/Philadelphia: John Benjamins, 27-49.

- **Halliday, Michael Alexander Kirkwood** (1966). "Lexis as a linguistic level." Charles Ernest Bazell *et al*. (eds) (1966). *In Memory of J.R. Firth*. London/Harlow: Longman, 148-162.

- — (1998). "Things and Relations: Regrammaticizing Experience as Technical Knowledge." Jonathan Webster (ed.) (2004). *The language of science*. London/New York: Continuum, 49-101.

- **Halliday, Michael Alexander Kirkwood and Ruqaiya Hasan** (1976). *Cohesion in English*. London: Longman.

- **Hausmann, Franz Josef** (1979). "Un dictionnaire des collocations est-il possible?" *Travaux de linguistique et de littérature* 17(1), 187-195.

- — (1989). "Le dictionnaire de collocations." Franz Josef Hausmann *et al.* (eds) (1989). *Wörterbücher: ein internationales Handbuch zur Lexicographie.* Vol. 1. Berlin/New-York: W. de Gruyter, 1010-1019.

- **Heid, Ulrich et Gerhard Freibott** (1991). "Collocations dans une base de données terminologiques et lexicales." *Meta : journal des traducteurs* 36(1), 77-91.

- **Hoey, Michael** (1991). *Patterns of Lexis in Text*. Oxford: OUP.

- — (2004). "Textual colligation: a special kind of lexical priming." *Language and Computers* 1(49), 171-94.

- — (2005). *Lexical priming: a new theory of words and language*. New York: Routledge.

- **Howarth, Peter Andrew** (1996). *Phraseology in English Academic Writing: Some Implications for Language Learning and Dictionary Making.* Tübingen: Max Niemeyer.

- **Hunston, Susan and Gill Francis** (2000). *Pattern Grammar: A corpus-driven approach to the lexical grammar of English*. Studies in Corpus Linguistics 4. Amsterdam/Philadelphia: John Benjamins.

- **Hyland, Ken** (2004) *Disciplinary discourses: social interactions in academic writing*. Michigan Classics Edition. Ann Arbor MI: The University of Michigan Press.

- **Kübler, Natalie and Mojca Pecman** (2012). "The ARTES bilingual LSP dictionary: from collocation to higher order phraseology." Sylviane Granger and Magali Paquot (eds) *Electronic lexicography*. Oxford: OUP, 186-208.

- **L'Homme, Marie-Claude et Isabelle Meynard** (1998). "Le point d'accès aux combinaisons lexicales spécialisées : présentation de deux modèles informatiques." *TTR : traduction, terminologie, rédaction* 11(1), 199-227.

- **Martinez, William** (2003). *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*. Thèse de doctorat. Université de la Sorbonne nouvelle – Paris.

- **Mel'čuk, Igor Aleksandrovič** (1993). "La phraséologie et son rôle dans l'enseignement / apprentissage d'une langue étrangère." *Etudes de Linguistique Appliquée* 92, 82-113.

- **Meynard, Isabelle** (1997). "Approche hypertextuelle via HTML pour un outil de consignation bilingue des combinaisons lexicales spécialisées." *Actes du Congrès international de terminologie,* San Sebastian (Espagne), 12-14 novembre 1997, San Sebastian : IVAP/UZEI, 675-689.

- **Partington, Alan** (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam/Philadelphia: John Benjamins.

- **Pavel, Silvia** (1993). "La phraséologie en langue de spécialité. Méthodologie de consignation dans les vocabulaires terminologiques." *Terminologies nouvelles* 10, 23-35.

- **Pecman, Mojca** (2007). "Approche onomasiologique de la langue scientifique générale." *Revue française de linguistique appliquée* 12(2), consacré au Lexique des écrits scientifiques, 79-96.

- — (2008). "Compilation, formalisation and presentation of bilingual phraseology: problems and possible solutions." Sylviane Granger and Fanny Meunier (eds) (2008). *Phraseology in language learning and teaching.* Amsterdam/Philadelphia: John Benjamins, 203-222.

- **Pecman, Mojca *et al*.** (2010). "Processing collocations in a terminological database based on a cross-disciplinary study of scientific texts." Sylviane Granger and Magali Paquot (eds) (2010). *eLexicography in the 21st Century: New Challenges, New Applications.* Proceedings of eLex 2009, Louvain-la-Neuve (Belgium), 22-24 October 2009. Louvain-la-Neuve: Cahiers du CENTAL, 249-262.

- **Pecman, Mojca and Natalie Kübler** (2011). "ARTES: an online lexical database for research and teaching in specialized translation and communication." *Proceedings from International Workshop on Lexical Resources (WoLeR) 2011 at ESSLLI*. August 1-5, 2011 – Ljubljana, Slovenia, 87-93. http://alpage.inria.fr/~sagot/woler2011/WoLeR2011/Program_&_Proceedings_files/WoLeR%202011%20-%20Pecman%20Ku%CC%88bler.pdf (consulted 22.06.2012).

- **Rastier, François** (1987). *Sémantique interprétative*. Paris: Presses Universitaires de France.

- — (1989). *Sens et textualité*. Paris: Hachette.

- **Siepmann, Dirk** (2006). "Collocation, Colligation and Encoding Dictionaries. Part II: Lexicographical Aspects." *International Journal of Lexicography* 19(1), 1-39.

- **Simpson-Vlach, Rita and Nick Ellis** (2010). "An Academic Formulas List: New Methods in Phraseology Research." *Applied Linguistics* 31(4), 487-512.

- **Sinclair, John McHardy** (1966). "Beginning the Study of Lexis." Charles Ernest Bazell *et al*. (eds) (1966). *In Memory of J.R. Firth*. London/Harlow: Longman, 410-430.

- — (1991). *Corpus, Concordance, Collocation*. Oxford: OUP.

- **Swales, John Malcolm** (1990). *Genre analysis: English in academic and research settings*. Cambridge Applied Linguistics Series. Cambridge: Cambridge University Press.

- **Tutin, Agnès** (2007). "Modélisation linguistique et annotation des collocations : application au lexique transdisciplinaire des écrits scientifiques." Svetla Koeva, Denis Maurel and Max Silberztein (eds) (2007). *Formaliser les langues avec l'ordinateur*. Besançon: Presses universitaires de Franche-Comté, 189-215.

- — (2008). "L'apport des corpus annotés pour l'élaboration semi-automatique d'une base de collocations de la langue scientifique générale." François Maniez *et al*. (eds) (2008). *Corpus et dictionnaires de langues de spécialité*. Grenoble: Presses Universitaires de Grenoble, 45-65.

- **Viprey, Jean-Marie** (2006). "Structure non-séquentielle des textes." *Langages* 163, 71-85.

- **Volanschi, Alexandra** (2008). *Étude et modélisation des phénomènes collocationnels : Implémentation dans un système d'aide à la rédaction en anglais scientifique*. Thèse de doctorat. Université Paris Diderot.

- **Williams, Geoffrey Clive.** (1999). *Les réseaux collocationnels dans la construction et l'exploitation d'un corpus dans le cadre d'une communauté de discours scientifique*, Thèse de doctorat. Université de Nantes.

- **Zinglé, Henri** (1994). "The ZStation workbench and the modelling of linguistic knowledge. Carlos Martin-Vide (ed.) (1994). *Current issues in mathematical linguistics.* Amsterdam/New York: North Holland, 423-432.

- — (1996). "ZART: un linguiciel d'aide à la rédaction scientifique et technique en langue étrangère." *Travaux du LILLA* 1, 111-113.

- — (1998). "ZTEXT: un outil pour l'analyse de corpus." *Travaux du LILLA* 3, 69-78.

- **Zinglé, Henri et Marie-Louise Brobeck-Zinglé** (2003). *Dictionnaire combinatoire du français. Expression, locutions et constructions*. Paris: La maison du dictionnaire.

Webographie

- "**Le site du projet ARTES**." http://www.eila.univ-paris-diderot.fr/artes (consulté le 18.06.2012).

- "**Interface d'accès au ressources de la base ARTES**." https://artes.eila.univ-paris-diderot.fr (consulté le 18.06.2012).

- "**La banque de données terminologiques et linguistiques du gouvernement du Canada**." http://www.termiumplus.gc.ca (consulté le 18.06.2012).

- "**Le grand dictionnaire terminologique de l'Office québécois de la langue française**." http://www.granddictionnaire.com (consulté le 18.06.2012).

- "**InterActive Terminology for Europe database**." http://iate.europa.eu (consulté le 18.06.2012).

- "**Dictionnaire fondamental de l'informatique et de l'internet**." http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi (consulté le 18.06.2012).

- "**Dictionnaire fondamental de l'environnement**." http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi (consulté le 18.06.2012).

- "**Le projet de Institute for Information Management à Cologne**." http://www.iim.fh-koeln.de/webterm/webtermsamm_e.htm (consulté le 18.06.2012).

- "**Projet et plateforme logicielle Textométrie**." http://textometrie.ens-lyon.fr (consulté le 18.06.2012).

Biography

Mojca Pecman is in charge of courses on terminology and lexical resources creation in the department of Applied Languages at Paris Diderot University. Her research focuses on terminology, phraseology, discourse analysis, LSPs, specialised translation, and corpus linguistics. She is currently carrying out a research project on the creation of terminological and phraseological resources which led to the design of the ARTES online multilingual and multidomain database with a view to improving translation-related research and studies. Contact: mpecman@eila.univ-paris-diderot.fr.