# Formulaire de soumission (document scientifique)/ Antragsformular

| | |
|---|---|
| Appel à projets franco-allemand en sciences humaines et sociales | Ausschreibung eines deutsch-französischen Programms in den Geistes- und Sozialwissenschaften |
| Programme non-thématique | Ohne thematische Vorgaben |
| 2015 | 2015 |
| Titre du projet | Titel des Projekts |
| **La phraséologie du roman** | **Phraseologie des Romans** |
| Acronyme | Kennwort |
| PHRASEOROM | PHRASEOROM |
| Nom du coordinateur français et institution(s) de rattachement | Namen des Projektleiters auf deutscher Seite und seiner Institution |
| Iva NOVAKOVA<br>LIDILEM, EA (609)<br>Université Stendhal, Grenoble 3, | Dirk SIEPMANN<br>Université d'Osnabrück, Institut d'Études anglophones et américaines |

## 1. FICHE D'IDENTITE DU PROJET

### 1.1. Partenariat et participants

Équipe française :

| Prénom | NOM | Statut | Rattache-ment | Coordonnées | Rôle dans le projet | Temps d'engagement |
|---|---|---|---|---|---|---|
| Iva | Novakova | Professeur | Lidilem E.A. 609 | Université Stendhal, Grenoble 3, BP 25 F-38040 Grenoble cedex Iva.Novakova@u-grenoble3.fr | Coordinatrice | 30 |
| Olivier | Kraif | MCF, HDR | Lidilem | Olivier.Kraif@u-grenoble3.fr | Participant direct au projet | 30 |
| Francis | Grossmann | Professeur | Lidilem | Francis.Grossman@u-grenoble3.fr | Participant direct au projet | 18 |
| Agnès | Tutin | Professeur | Lidilem | Agnes.Tutin@u-grenoble3.fr | Participant direct au projet | 12 |
| Julie | Sorba | Docteur en SCL, PRAG | Lidilem | Julie.Sorba@u-grenoble3.fr | Participant direct au projet | 12 |
| Julien | Piat | MCF | LITT&ARTS E.A. 7355 Equipe Traverses 19-21 | Julien.Piat@u-grenoble3.fr | Participant direct au projet | 18 |
| Laetitia | Gonon | MCF | LITT&ARTS E.A. 7355 Equipe Traverses 19-21 | Laetitia.Gonon@u-grenoble3.fr | Participant direct au projet | 18 |
| Dominique | Legallois | MCF, HDR | Crisco EA 4255 | Université de Caen, Campus 3 Boulevard Yitzhak Rabin , 14123 IFS dominique.legallois@unicaen.fr | Participant direct au projet | 18 |

Autres contributeurs (de l'équipe française) :

| Prénom | NOM | Statut | Rattachement(s) | Coordonnées |
|---|---|---|---|---|
| Gilles | Philippe | Professeur ordinaire | Université de Lausanne, Centre de recherche en langues et littératures comparées (CLE) | Unil-Dorigny, Bâtiment Anthropole, CH-1015 Lausanne<br><br>gilles.philippe@unil.ch |
| François | Maniez | Professeur | Université Lyon 2, | 6 rue Pasteur, 69365 Lyon cedex 7 |

| | | | | |
|---|---|---|---|---|
| | | | Directeur du Centre de recherche en terminologie et en traduction EA 4162 | francois.maniez@univ-lyon2.fr |
| Teresa | Muryn | Professeur | Institut pédagogique de Cracovie, Pologne Directrice de la Chaire de linguistique romane ; Directrice adjointe de L'Institut de Lettres et de Langues (ILLM) | Instytut Neofilologii<br>Uniwerstet Pedagogiczny<br>Ul. Podchorazych 2<br>30-084 Krakow<br><br>teresa.muryn@gmail.com |
| Małgorzata, | Niziolek | Maître de conférences | Institut pédagogique de Cracovie, Pologne, ILLM, Chaire de linguistique romane | Instytut Neofilologii<br>Uniwerstet Pedagogiczny<br>Ul. Podchorazych 2<br>30-084 Krakow<br><br>mniziolek1@gmail.com |

Équipe allemande :

| **Prénom** | **NOM** | **Statut** | **Rattachement(s)** | **Coordonnées** | **Rôle dans le projet** | **Temps d'engagement** |
|---|---|---|---|---|---|---|
| Dirk | Siepmann | Professeur | Université d'Osnabrück, Institut d'Études anglophones et américaines | Universität Osnabrück, Neuer Graben 40, D-49069 Osnabrück, dirk.siepmann@uos.de | Coordinateur | |
| Ludwig | Fesenmeier | Professeur | Université d'Erlangen | Bismarckstr1, D-91054 Erlangen ludwig.fesenmeier@fau.de | Participant direct au projet | |
| Marion | Gymnich | Professeur | Université de Bonn, Département d'Études anglophones, américaines et celtiques | Universität Bonn, Regina-Pacis-Weg 5 D-53113 Bonn, mgymnich@uni-bonn.de | Participant direct au projet | |
| Sascha | Diwersy | Docteur | Université de Cologne | Universität zu Köln, Albertus-Magnus-Platz, D-50923 Köln sascha.diwersy@uni-koeln.de | Participant direct au projet | |
| N .N. | | Doctorant/ Habilitant | Université d'Osnabrück | | Participant direct au projet | |

| | | | | | | |
|---|---|---|---|---|---|---|
| N .N. | | Doctorant/ Habilitant | Université de Bonn | | Participant direct au projet | |
| N.N. | | Doctorant/ Habilitant | Université d'Erlangen | | Participant direct au projet | |

Autres contributeurs (de l'équipe allemande) :

| **Prénom** | **NOM** | **Statut** | **Rattachement(s)** | **Coordonnées** |
|---|---|---|---|---|
| John Desmond | Gallagher | Chargé de cours | Université de Münster | Erlenallee 3, 48155 Münster, jdgallagher@gmx.de |
| Peter | Blumenthal | Professeur | Université de Cologne | Universität zu Köln, Albertus-Magnus-Platz, D-50923 Köln, peter.blumenthal@uni-koeln.de |

### 1.2. Titre du projet et acronyme

La phraséologie du roman PHRASEOROM

### 1.3. Discipline et domaine de recherche

Linguistique ; Phraséologie ; Linguistique de corpus ; Traitement automatique du langage (TAL) ; Études littéraires et stylistique ; Linguistique textuelle ; Linguistique contrastive ; Traductologie

### 1.4. Durée du projet

36 mois

### 1.5. Résumé

Le principal objectif de ce projet est d'élaborer, dans une démarche inductive *corpus-driven*, une typologie structurelle et fonctionnelle des constructions lexico-syntaxiques spécifiques (CLS) au discours romanesque francophone, anglophone et germanophone du XX$^{e}$ siècle, le roman constituant le genre littéraire qui touche le lectorat le plus large. Sur la base de cette typologie, on procédera à deux types de comparaison :

a) entre littérature et paralittérature (angl. popular literature, all. Trivialliteratur ; science-fiction ; roman policier ; roman sentimental) ;
b) entre les pratiques stylistiques observables dans des traditions littéraires de pays différents (Royaume Uni, France, Allemagne).

Dans un premier temps, nous effectuerons des calculs statistiques qui permettront d'établir les récurrences significatives des constructions lexico-syntaxiques au sein des textes littéraires par rapport à un corpus de contraste (journalistique, scientifiques). Nous chercherons ensuite à établir, sur de grands corpus textuels, dans quelle mesure ces unités lexicales étendues jouent-elles un rôle dans la construction du texte littéraire et proposerons une typologie de ces unités. L'analyse linguistique des données sur les plans sémantique, syntaxique et discursif sera articulée à une analyse stylistique au sein de différents genres romanesques dans un but comparatif. Il s'agit de jeter les fondements d'un « lexique-grammaire » des constructions spécifiques au roman, avec des retombées en linguistique et en stylistique contrastives, ainsi qu'en traductologie.

Il s'agit d'un projet interdisciplinaire au croisement de la linguistique et des études littéraires et, en particulier, de la phraséologie, de la stylistique, de la théorie des genres, de la linguistique de corpus et du traitement automatique du langage (TAL). Par son objet de recherche (les phraséologismes du roman) et sa méthodologie (celle de la linguistique de corpus outillée), le projet relève du domaine des Humanités numériques en Sciences humaines et sociales.

## 2. ÉTAT DE LA RECHERCHE, TRAVAUX ANTERIEURS

### 2.1. État de la recherche

#### 2.1.1. Contexte scientifique : convergence entre stylistique littéraire et linguistique

À l'heure actuelle, la phraséologie suscite une attention grandissante au sein de deux courants de recherche traditionnellement séparés : la linguistique outillée et l'analyse stylistique du texte littéraire.

D'une part, la linguistique outillée (cf. pour ce terme Habert 2005) a développé des procédés automatiques pour l'extraction des items lexicaux qui favorisent une meilleure approche de la polylexicalité. À cela s'ajoute une réflexion plus large menée en phraséologie sur le fonctionnement des phraséologismes (cf. Europhras 2014). L'accent est mis sur la nécessité d'appliquer une approche globale qui intègre les niveaux lexico-sémantique, syntaxique et discursif dans le traitement des phraséologismes (cf. *Rencontres phraséologiques*, 2013, Grenoble).

D'autre part, des développements récents de l'analyse stylistique ont mis en évidence l'existence de patrons stylistiques, outil fécond pour penser non seulement l'historicité des formes littéraires mais encore celle de leur réception – et donc celle de leur lecture (Maingueneau & Philippe 1999, Philippe et Piat 2009, Vaudrey-Luigi 2013).

Par ailleurs, plusieurs colloques internationaux sont consacrés, pendant la seule année 2015, aux différentes approches à privilégier dans la *Grammaire des genres et des styles (*Journée ConsCiLa , janvier 2015, Paris), au croisement et aux frontières de la *Grammaire et* [des] *textes* (GRATO, juillet 2015, Lisbonne), au renouvellement épistémologique et heuristique des théories textuelles et discursives dans l'espace européen (*Texte et discours en confrontation dans l'espace européen,* Metz, septembre 2015), à l'articulation entre *Marqueurs et structures dans la (re)construction du sens* (novembre 2015, Paris-Nanterre), etc.

Le contexte scientifique actuel offre la possibilité de mettre en synergie les forces de ces deux approches qui se fécondent mutuellement. Ainsi, les questions de recherches et les résultats de type plutôt exploratoire qu'on rencontre en stylistique doivent faire l'objet d'une validation statistiquement fiable sur grands corpus. À l'inverse, l'expertise littéraire est indispensable pour confirmer la pertinence des phénomènes identifiés par l'analyse automatique de textes. Tout cela rejoint nos objectifs et reste à faire sur une grande échelle dans le cadre d'un projet international.

#### 2.1.2 Approches linguistiques de la littérature

Il existe de nombreux travaux en stylistique (Barthes 1966, Greimas 1972/1982, Leech & Short 2007), en stylométrie (Magri-Mourgues 2006), en stylistique de corpus (Stubbs 2005, Fischer-Starcke 2010, Mahlberg 2013) et en textométrie (Guiraud 1954, Brunet 1981) qui traitent de manière approfondie des procédés de style et des spécificités lexicales et grammaticales chez différents auteurs (p.ex. Flaubert, Proust, Dickens, Jane Austen) ou qui établissent des schémas stéréotypiques récurrents, p. ex. pour le roman policier (Todorov 1980, Marion 2009,

Lits 2011), dans le but de dresser une typologie des textes littéraires. Cependant, à l'exception de Legallois (2012) et de Siepmann (2015b, 2016), les constructions lexico-syntaxiques (CLS) utilisées en littérature n'ont pas fait l'objet d'études systématiques d'un point de vue strictement linguistique (à la fois sur les plans sémantique, syntaxique et discursif) et encore moins sur des corpus de fort volume et avec les outils du TAL.

Nombre de linguistes s'accordent d'ailleurs à estimer que les textes littéraires ne méritent pas de statut spécial, dans le sens où les caractéristiques formelles souvent désignées comme littéraires, telles que les métaphores ou l'ironie, apparaissent également dans d'autres types de discours (Burton & Carter 2006 : 273). Ce positionnement est à l'origine des tentatives visant à définir la spécificité du langage littéraire par une approche fonctionnelle, selon des critères tels que la « duplicité » (Scholes 1982) ou selon différents facteurs intervenant dans le processus de communication (p.ex. la distinction entre auteur et narrateur). Dans cette optique, l'impression subjective de « littérarité » (Jakobson 1960) qui émane même d'un bref extrait de roman ne serait qu'une sorte d'épiphénomène aléatoire venant se greffer sur le travail de l'écrivain. Or, cette affirmation d'une différence fonctionnelle sans corrélat formel paraît sans précédent en linguistique. Elle tient probablement au fait que l'attention se soit centrée sur le style de certains auteurs ou de certains textes plutôt que sur l'étude de grands corpus littéraires.

Notre recherche repose donc sur l'hypothèse que le langage romanesque, aussi bien dans son ensemble que dans ses genres spécifiques, se caractérise par la surreprésentation statistiquement significative de certains phénomènes linguistiques (lexèmes, collocations, colligations, schémas actanciels), lesquels jouent un rôle non négligeable dans la construction littéraire du texte. Ceci n'exclut nullement qu'il puisse y avoir rupture avec des patrons individuels, comme le postulait déjà le formalisme russe, mais, en règle générale, ces ruptures représentent soit des variantes plus élaborées de certains stéréotypes référentiels (*a silence fell → a blanket of strained silence instantly descended*), soit des variantes qui se démarquent d'une façon ou d'une autre de ces stéréotypes (appelés « stéréotypes du deuxième degré », Dufays (2010 : 239) ; par ex. *an economically dressed nymphet* au lieu de *a scantily dressed young girl* (Siepmann 2015a : 106-107).

L'hypothèse que l'on vient d'émettre a déjà fait l'objet d'une amorce de validation. D'un point de vue très général, on notera l'étude pionnière sur l'anglais de Stubbs & Barth (2003), qui démontrent que « les types de textes se distinguent par les séquences lexicales et grammaticales qu'ils renferment » (« text types are distinguished by lexical and grammatical patterns » [p. 79]). Force est pourtant de constater que l'étude en question, en plus d'être fondée sur un corpus de taille modeste, se focalise uniquement sur les 200 segments répétés les plus fréquents dans chaque type de texte, ce qui exclut plus ou moins les unités « lexicalement riches » qui influent grandement sur les perceptions du lecteur.

Une deuxième source de renseignements sur les spécificités du discours littéraire anglais est la grammaire descriptive de Biber *et al.* (1999), qui identifie un certain nombre de propriétés grammaticales spécifiquement littéraires (les doubles génitifs,les constructions démonstratives du type « that bloody car of mine » ou les propositions existentielles en « there » avec des verbes autres que *be*). Sont significativement non spécifiques la postmodification par une relative introduite par un participe passé et les appositions.

Si une telle cartographie des propriétés grammaticales est précieuse, elle ne permet pas pour autant d'observer les combinaisons spécifiques de certaines catégories grammaticales. Autrement dit, fait défaut le répertoire de séquences syntagmatiques qu'un genre préfère au détriment d'autres constructions. Pour combler cette lacune dans le

domaine du français, Legallois (2012) emploie la méthode dite des motifs séquentiels (voir 2.1.4.) qui permet de rendre compte des séquences syntagmatiques d'un texte et de caractériser les spécificités lexico-grammaticales d'un genre ou d'un auteur. Ainsi, Legallois met en évidence chez huit auteurs différents du XIX$^{e}$ s. le motif « le N1 du N2 qui V dans », qui sert de cadre à des réalisations lexicales groupées autour des isotopies suivantes : vent {*vent, brise, bise*}, nuit {*nuit, soir*}, mort {*mourir, gémir, pleurer, tomber, se briser, abîme*}, feuillage {*feuillage, feuille, branches, bois*}. À la lumière d'exemples comme celui-ci, on peut penser que les motifs constituent des unités discursives liées à un certain type de phraséologie, et qu'ils permettent de repérer des relations intra- et intertextuelles intéressantes pour la recherche en stylistique ainsi qu'en théorie et histoire des genres. On retrouve ici la méthodologie des *concgrams* mise au point par Greaves (2005), mais appliquée à des constructions grammaticales.

L'étude de Siepmann (à paraître, 2016), dans le prolongement de ces travaux, soumet un corpus de romans anglophones de 160 millions de mots à trois types d'analyses issues d'une démarche *corpus-driven* : a) étude des mots-clefs (*key word analysis*), b) étude des segments répétés (Lebart & Salem 1988) ou « paquets lexicaux » (*lexical bundles*, cf. Biber *et al.* 1999), c) analyse détaillée de trois mots-clefs de type différent (*thought*, *sun*, *jerk*). Siepmann identifie douze classes principales de mots-clefs anglais (p.ex. les verbes de mouvement), qui se révèlent identiques dans un corpus parallèle francophone. En revanche, la fréquence de certains équivalents de traduction (les noms *seins* et *breasts*, le verbe *nod* et ses équivalents français *faire un signe de la tête, acquiescer de la tête, hocher le front, , dodeliner de la tête*, etc.) peut varier sensiblement d'une langue à l'autre (cf. Siepmann 2015b). Il s'avère en outre que la presque totalité des paquets lexicaux les plus fréquents formés à partir de *like* est basée sur les mots-clefs (p.ex. *sun was like a*, *room was like a*, *voice was like a*). Enfin, l'étude approfondie de trois mots-clefs (*thought*, *sun*, *jerk*) met en évidence un nombre important de séquences lexico-grammaticales spécifiques aux romans, dont beaucoup peuvent être décrites comme des « unités lexicales étendues » au sens de Sinclair (p.ex. le marqueur de transition « génitif + *thoughts were on* + GN »). Ainsi, l'analyse du nom concret *sun* montre que les structures anglaises du type *the sun trickled through the trees* donnent lieu à de multiples variations par simple manipulation du slot verbal (cf. *slant/pour/lick/seep/… through*). Rien de tel en français, où l'expression du même contenu sémantique s'avère plus complexe (p.ex. *des gouttes de soleil pénétraient à travers le feuillage*).

### 2.1.3. Approches de l'idiomaticité et de la phraséologie

Si, comme on vient de le voir, il existe peu de travaux sur la nature lexico-grammaticale des textes littéraires, de nombreux chercheurs se sont penchés sur les combinaisons idiomatiques dans d'autres genres tels que les textes journalistiques et scientifiques (Sinclair 1991, 2004, Hunston & Francis 2000, Hoey 2005, de Beaugrande 2005). L'idiomaticité se manifeste dans les textes par tout un éventail de phraséologismes qui échappent, pour l'instant, à une définition complète et consensuelle : « extended units of meaning » (Sinclair 2004), « constructions » (Goldberg 1995), « collostructions » (Stefanowitsch & Gries 2003), « collocations » (Hausmann 1979, Mel'čuk *et al.* 1995, Siepmann 2005, Tutin 2010), « lexical bundles » (Biber et al. 1999), « motifs séquentiels » (S. Quiniou et *al*. 2012), « les combinaisons de mots usuelles » (Steyer 2013).

On observe néanmoins une convergence croissante de ces différentes appellations et approches, qui ont en commun d'abandonner la distinction entre une grammaire composée

de règles et un lexique composé de mots et de locutions. Ainsi, le courant néofirthien, dont le développement le plus abouti est sans doute la *Lexical Priming Theory* de Hoey (2005), postule un lexique-grammaire fait de combinaisons de nature grammaticale („colligations": ex. GN + *to be* + *about* + V-*ing*) et de combinaisons de nature lexicale („collocations": *clear motorway*). La grammaire constructionnelle, elle, appréhende la langue comme un inventaire d'appariements conventionalisés entre forme et fonction, qui s'étendent sur un continuum allant du lexique aux structures grammaticales, en passant par les séquences idiomatiques (p.ex. Goldberg 1995, Fillmore *et al*. 1998, Croft 2001). Il y a une parenté évidente entre le contextualisme et certains courants de la grammaire constructionnelle tels que les « collostructions » de Stefanowitsch & Gries (2003). Alors que les premiers partent de constructions à caractère général telles que la construction ditransitive (V + GN + GN) pour voir quel lexique s'y associe, les seconds empruntent le chemin inverse, en partant des lexèmes individuels (p.ex. *give*) pour arriver au schème actanciel abstrait. Enfin, des approches sociolinguistiques et pragmatiques telles que celle de Feilke (1994, 1996, 2003) font appel à des théories sociologiques pour mettre en évidence la constitution sociale de l'usage idiomatique des langues. Ces approches font du phénomène phraséologique, jusqu'ici marginalisé dans la théorie linguistique, le principe central du fonctionnement de la langue, caractérisant la compétence langagière comme une « compétence de sens commun » qui se fonde sur des sélections possibles et confirmées par les acteurs sociaux.

La forte convergence de ces différentes approches fait que le regard des chercheurs se déplace des séquences figées de la phraséologie traditionnelle (formules routinisées, proverbes, collocations binaires (Hausmann 1979, Mel'čuk *et al.* 1995, Tutin 2010) vers toutes sortes de « pragmatèmes » (Feilke 1996) et d'unités lexicales « étendues » (Sinclair 2004). Deux notions particulièrement prometteuses à ce point de vue sont celles de « cadres collocationnels » (Renouf & Sinclair 1991), « motifs » (voir *supra* ; Legallois 2006, Longrée & Mellet 2013) ou « probabèmes » (Herbst & Klotz 2003), ces derniers étant des séquences polylexicales que les locuteurs emploient avec une certaine probabilité (p.ex. *il n'avait pas de mots assez durs pour INF* (Hausmann 2007: 136; cf. ?*il avait des mots durs pour INF*).

C'est dans la théorie du *lexical priming* de Hoey (2005) que l'on retrouve la présentation la plus englobante du phénomène combinatoire. Il fait appel à la notion de « nids » collocationnels, dont le sens n'est pas compositionnel (*say a word -> say a word against -> won't say a word against* [Hoey 2005: 11]). En outre, il complète la description des relations lexico-syntaxiques « locales », c'est-à-dire internes à la phrase, par l'analyse des relations lexico-syntaxiques dans des unités textuelles plus vastes. Il n'en reste pas moins que le domaine des nids collocationnels et colligationnels n'a pas fait l'objet d'études détaillées sur de grands corpus et encore moins dans une perspective interlinguistique. Dans le cadre d'EMOLEX, nous avons néanmoins mis au point des procédés en TAL permettant d'aller au-delà des combinaisons de deux mots.

### 2.1.4. L'application du TAL à l'extraction de constructions lexico-syntaxiques

Des applications récentes en TAL permettent désormais d'aller au-delà du traitement traditionnel de simples *n-grammes* de mots-formes. Une approche prometteuse dans ce domaine est adoptée par Quiniou *et al.* (2012) et Legallois (2012), qui exploitent les séquences en surface au moyen d'une technique bien connue en fouille de données : l'extraction de motifs séquentiels. Dans ce cadre, les motifs séquentiels doivent être vus comme des suites d'unités continues ou discontinues ; ils sont séquentiels, c'est-à-dire que les unités sont

ordonnées. Ces motifs peuvent être des *items* (un item représente une seule information, p.ex.. la forme d'un mot) ou des *itemsets* (combinaison de plusieurs types d'information, p.ex., des formes de mots, des lemmes ou des catégories morpho-syntaxiques). On peut considérer que les motifs sont des extensions des *segments répétés* ou *n-grammes* : alors que ces derniers constituent des suites continues de mots (p.ex. *était une fois, don't know what*), les motifs, outre le fait qu'ils peuvent s'appliquer à plusieurs niveaux de généralisation, supportent des *gaps* (empans) programmables entre les unités. Ces motifs s'inscrivent donc dans le prolongement direct des « phrase-frames » proposées par Fletcher (2007) (avec cette différence que les phrase-frames ne comportent qu'un seul « trou », p.ex. * *was like a*). Ils se rapprochent aussi des *concgrams* proposés par Greaves (2005), terme sous lequel on désigne toutes les cooccurrences de plusieurs mots, indépendamment de leur morphologie ou de leur position dans les énoncés. Par exemple, selon Legallois (2012), le motif suivant est spécifique du style de Zola (selon le calcul des spécificités utilisé en textométrie, et en comparaison avec un corpus de romans du XIX$^{e}$ s.) : ***il V le NC de DETPOSS NC ADJ*** (p.ex. « Elle ne parla plus, elle s'abattit près du brancard, dont *elle écarta les toiles de ses mains tremblantes* » ou « Alors, ils cessèrent de rire, penchés au-dessus de la Bible antique, dont *elle tournait les pages, de ses doigts minces*. »).

Une autre approche innovante qui va au-delà de la simple séquence de mots en surface est proposée par Kraif et Diwersy (2012). Cette approche bénéficie du développement récent de parseurs performants (XIP, Malt Parser, Connexor, IDP Parser, Mate Tools, Talismane, FIPS, DeSR) permettant d'extraire automatiquement, avec une bonne précision, des relations syntaxiques dans les corpus de grande dimension. C'est un tout nouveau champ d'exploration qui s'ouvre pour l'étude de la phraséologie. En effet, les cooccurrences syntaxiques présentent un réel intérêt en termes de bruit et de silence, car la fenêtre au sein de laquelle s'effectue la prise en compte des cooccurrences n'est pas arbitraire (cf. Evert 2008). Dans la continuité de Seretan *et al.* (2003), Charest *et al.* (2010) se sont basés sur ce type de cooccurrence pour extraire les collocations du dictionnaire Antidote RX, en élargissant le champ à des expressions « multilexémiques » pouvant comporter plus de deux mots.

Appliquant une technique d'extraction itérative, basée sur la notion de pivot complexe, Kraif et Diwersy mettent en place une méthode d'extraction à la demande, sans précalcul, des expressions récurrentes de longueur *n* autour d'un pivot donné. Ces expressions récurrentes dépassent le simple cadre des segments répétés ou paquets lexicaux. Elles représentent de véritables sous-arbres syntaxiques récurrents, susceptibles de se réaliser, de différentes manières, en surface dans les textes, comme le montre l'exemple de la figure 1 (extrait du module EmoConc de l'EmoBase h.ttp://emolex.u-grenoble3.fr/emoBase/:

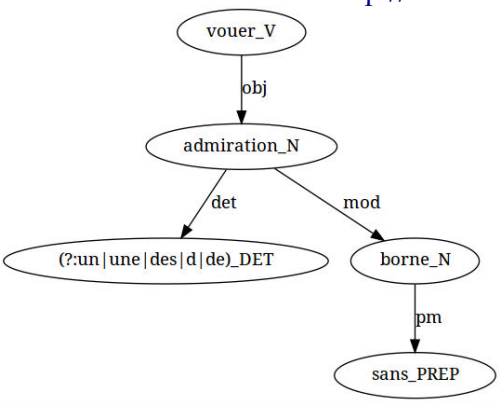


Figure 1: Extraction automatique du sous-arbre correspondant à « vouer une admiration sans bornes »

### 2.1.5. Stylistique et théorie des genres

Le travail sur la langue littéraire est un domaine de recherche récent dans le domaine stylistique : G. Philippe et J. Piat ont montré que, depuis l'autonomisation du champ littéraire en France, autour de 1850 (Bourdieu 1992, Vaillant 2003), une langue *autre* avait émergé en contexte littéraire. En prenant en compte le discours des écrivains sur leurs pratiques, en comparant génération par génération ces mêmes pratiques, en repérant des récurrences historiquement pertinentes (ainsi des tours du type *la propreté des vitres* en lieu et place de *des vitres propres* à la fin du XIX$^{e}$ s. ou de l'abondance des incidentes métadiscursives autour de 1960), en observant le discours des critiques et des grammairiens sur la littérature, c'est bien une langue spécifiquement littéraire qui s'est progressivement construite – comme le soulignent tant l'exclusion des exemples tirés des auteurs hors des grammaires (Philippe 2002) que, dans les années 1920, les querelles autour du « style » de Flaubert (Philippe 2004) ou du vaudois Ramuz.

La mise en évidence de patrons stylistiques est l'un des principaux bénéfices de cette recherche (cf. 2.1.1.). Définies comme série de traits co-occurrents construisant de manière conventionnelle (donc historique) tel ou tel effet de texte, ces configurations offrent un terrain de recherche interdisciplinaire extrêmement fécond en ouvrant de nouvelles perspectives sur l'historicité des formes littéraires et la catégorisation des genres sur une base linguistique. J. Piat (2011) a ainsi montré que l'écriture du Nouveau Roman des années 1950 reposait sur l'activation d'un patron phénoménologique : l'association d'un travail sur la longueur des phrases graphiques (étirées ou réduites à l'extrême) à divers phénomènes de discontinuité syntactico-énonciative vise à représenter différentes opérations mentales.

Philippe & Piat (2009) ont retracé, à partir d'une série d'études générationnelles, quels patrons stylistiques se configurent et dominent la prose française à tel ou tel moment de son histoire, ce qui leur permet de réfléchir au style comme pratique historiquement partagée entre auteurs et de cartographier le champ littéraire en définissant des pôles où la pratique de certains écrivains est congruente, quand elle s'oppose à celle d'autres auteurs (cf. p.ex. Vaudrey 2013).

On a également pu mettre en évidence des collocations lexicales stéréotypées, qui sont particulièrement fréquentes dans certains genres tels que les romans-feuilletons ; par exemple, dans ceux qui mettent en scène un vengeur, on va trouver de façon privilégiée la suite « GN désignant le bras ou la main » + *d'acier (le Chourineur sentit se tendre des nerfs et des muscles d'acier* [E. Sue]). Le roman-feuilleton emprunte à l'époque beaucoup de ses thèmes et de ses recettes d'écriture au fait divers ; ces deux genres de discours donnent d'ailleurs lieu, dans les années 1860, au roman policier. On considère ainsi que *L'Affaire Lerouge* d'É. Gaboriau est en 1866 le premier roman policier français. De même, le roman-feuilleton et le roman policier ont pu être considérés comme des sous-genres de la *paralittérature*. Cette catégorisation demande cependant à être interrogée, tant dans sa pertinence qu'en raison de l'extrême hétérogénéité (stylistique, littéraire) de ce qu'on appelle le genre du roman-feuilleton et le genre du roman policier.

Pour ce qui est de la théorie des genres, on constate que la littérature spécialisée existante tend à opérer une distinction assez nette entre littérature et paralittérature, faisant appel à des critères de différenciation basés exclusivement sur le contenu des genres en question, tels que par ex. la description d'états émotionnels dans les romans sentimentaux (cf. entre autres Gymnich, Neumann et Nünning 2007, Zymner 2003, Frow 2006, Duff 2000 Monte & Philippe 2014 sur les genres textuels). Cette approche par le contenu demande à

être complétée par une approche linguistique, ce qui ouvrira des perspectives nouvelles à la théorie des genres (cf.aussi Beauvisage 2001, Rastier, 2011). Idéalement, on aboutira à des correspondances entre le contenu littéraire et son expression linguistique. Les résultats obtenus sur les stéréotypes de différents genres permettront de réévaluer et réinterpréter certains courants littéraires et, par la suite, des auteurs et des œuvres individuelles.

### 2.1.6. Linguistique contrastive et traduction littéraire

En dépit d'une coopération croissante entre linguistes contrastivistes et traductologues, dont témoignent les priorités thématiques d'un certain nombre de revues spécialisées (p.ex. *Languages in Contrast*) et l'organisation régulière de conférences (tel que *Using Corpora in Contrastive and Translation Studies* [UCCTS]), la littérature qui explore l'interdépendance entre la linguistique contrastive et la traduction littéraire reste très mince. Plus particulièrement, il n'existe guère d'études récentes sur les ressemblances et dissemblances systématiques entre les usages littéraires des auteurs francophones, anglophones et germanophones. Or, la non prise en compte des ressemblances et dissemblances, en particulier dans le domaine des phraséologismes, peut donner lieu à des interprétations incorrectes de données traductionnelles de la part des traductologues (Granger *et al.* 2003 : 25) et constitue une source fréquente d'erreurs dans la pratique de la traduction. En outre, les résultats obtenus par les linguistes contrastivistes permettent d'apprécier plus aisément les équivalences proposées par les traducteurs littéraires.

Notre hypothèse est qu'on peut mettre en évidence des équivalences interlinguistiques systématiques entre phraséologismes littéraires (p.ex. *X interrupted his thoughts – X le tira de ses réflexions*), dont certains seraient spécifiques à des genres particuliers. Ceci ouvre deux pistes de recherche. Premièrement, il sera intéressant de voir dans quelle mesure des écrivains appartenant à des communautés linguistiques distinctes créent des combinaisons syntagmatiques analogues ou divergentes lorsqu'ils tentent d'exprimer des contenus identiques. Outre les correspondances entre les séquences récurrentes du type *X interrupted his thoughts*, on peut établir des équivalences *a priori* surprenantes dans des romans de langue différente : par ex. entre des syntagmes de faible fréquence tels que *un jour pluvieux se déversait dans la chambre* (H. Troyat) et *a sunny rainy light filled the room* (I. Murdoch) (cf. Gallagher 2007). D'un autre côté, il existe des configurations spécifiques à chaque langue, telles que les unités lexicales étendues fondées sur le nom anglais *sun* et les verbes à particule servant à décrire le passage de la lumière dans un endroit donné (*pour*, *slant*, *trickle* [*through the trees*]), qui donnent lieu à des variations aspectuelles multiples ; les configurations prototypiques équivalentes en français, quant à elles, sont atténuatives ou neutres du point de vue sémantique (*le soleil filtrait / glissait entre les [nom de l'arbre]* etc.) (Siepmann 2015b). Les répercussions sur la théorie de la traduction et sur l'activité traduisante sont non négligeables, dans la mesure où le traducteur hésitera entre des traductions normalisantes (*le soleil filtrait à travers le feuillage*), mais sémantiquement déficientes, et des traductions pleines, qui demanderont souvent des transpositions complexes. Or, les théories actuelles sur les divergences interlinguistiques de lexicalisation (formulées indépendamment par la stylistique comparée et par Talmy 2000) en rendent compte mais de façon incomplète.

Ces analyses permettront de mieux cerner un certain nombre de phénomènes discutés depuis longtemps en stylistique comparée (Malblanc 1961, Vinay & Darbelnet 1958, Truffaut 1983, Grünbeck 1976, François 1989, Chuquet & Paillard 1987, Ballard 1992) et de

remédier au peu d'attention que celle-ci porte à la question du genre et à l'historicité des formes linguistiques et littéraires. Tout se passe comme si les divergences interlinguistiques mises en évidence dans certains types de texte étaient valables pour la langue prise dans sa globalité et à n'importe quel moment de son histoire. Ce parti pris implique l'occultation de certains phénomènes propres à la langue littéraire, tels que les patrons stylistiques précédemment évoqués ou les collocations spécifiquement littéraires (p.ex. *nod in agreement*, *jerk awake*, cf. Siepmann 2016).

### 2.2. Travaux antérieurs

#### 2.2.1. L'équipe française

Les enseignants-chercheurs du LIDILEM participant à ce projet travaillent depuis plusieurs années en phraséologie (cf. entre autres Tutin 2013). Les travaux réalisés dans le cadre du projet ANR/DFG Emolex (www.emolex.eu) ont débouché sur la mise en place d'un modèle fonctionnel d'analyse des collocations d'émotion (Novakova & Melnikova 2013, Novakova 2015). Des pistes théoriques et méthodologiques ont été dégagées pour l'analyse discursive et textuelle de ces collocations dans les textes journalistiques et littéraires (Novakova & Sorba 2013a et b, 2014a et b). D'autres travaux importants en phraséologie du texte scientifique ont été menés dans le cadre du projet ANR Scientext http://scientext.msh-alpes.fr . Un ouvrage récemment paru (Tutin & Grossmann 2014) synthétise les recherches effectuées sur les phraséologismes liés au positionnement de l'auteur scientifique par rapport à ses devanciers et à ses pairs. Le repérage déjà effectué de différents types de routines discursives dans le discours scientifique représente un acquis important pour l'approche contrastive inter-genres qui sera appliquée à l'étude du discours romanesque. Dans le cadre du programme TAL, notre équipe a travaillé sur de nombreux projets impliquant de vastes corpus multilingues (projets Carmel, Scientext, Emolex), et a ainsi constitué une solide expérience dans la mise en œuvre de chaînes de traitement complexes (balisage, annotation syntaxique, alignement, indexation) et dans le développement d'outils et d'interface pour l'interrogation et l'extraction de données textuelles (Kraif 2008, Kraif & Diwersy 2012, 2014). Les travaux sur les corpus bilingues et parallèles ont débouché sur des réflexions méthodologiques menées dans le cadre de la linguistique de corpus (Melnikova, Novakova, Kraif 2009).

L'expertise de D. Legallois, enseignant-chercheur du Crisco (Université de Caen) sur les motifs séquentiels dans les romans du XIX$^{e}$ s., sur la fouille de données pour la stylistique (Legallois 2006, 2012) sera d'une grande utilité pour les aspects linguistiques et informatiques de ce projet. D. Legallois a publié en commun avec A. Tutin des travaux en phraséologie (cf. entre autres Legallois & Tutin 2013).

Le projet s'inscrit également dans la continuité des travaux menés en stylistique et poétique historique par les chercheurs de l'EA Litt&Arts (Traverses 19-21). Les travaux de J. Piat sur la langue littéraire (Philippe & Piat : 2009 ; Piat : 2006, 2011, 2013) ont posé des bases théoriques et méthodologiques sur la notion de patrons stylistiques (cf. 2.1.5). Les travaux de L. Gonon (2010, 2012) portent sur les figements et les interdiscours médicaux, policiers et littéraires dans les faits divers du XIX$^{e}$ s. ; elle s'intéresse également aux collocations, phraséologismes et lieux communs dans les romans-feuilletons.

#### 2.2.2. L'équipe allemande

***Osnabrück***

Les travaux menés par D. Siepmann et ses collaborateurs s'inscrivent dans le domaine de la phraséologie et de la linguistique contrastive. D. Siepmann (2005a) établit une taxonomie des

unités polylexicales à valeur discursive ou « marqueurs discursifs polylexicaux » et propose une analyse contrastive détaillée (anglais, français, allemand) de leur fonctionnement discursif. En suivant une démarche contextualiste inductive, Siepmann (2002, 2004, 2005b) problématise et affine la notion de « collocation ». Siepmann (2011) étend le principe d'idiomaticité de Sinclair (1991) en postulant la nécessité d'un principe de créativité. Dans le cadre du projet Emolex, Siepmann a publié deux études (2014b, c) sur les divergences sémantiques entre les collocations qu'admettent les noms d'émotion en français, en anglais et en allemand. B. Kern (2014), en collaboration avec A. Grutschus de l'équipe colonaise, a livré deux études contrastives sur la polarité du lexique de l'affect et sur le comportement discursif des noms d'émotion. Siepmann (2012) synthétise ses recherches effectuées sur les phraséologismes dans les textes scientifiques anglais dans un manuel à l'intention des auteurs scientifiques de langue maternelle allemande. Siepmann (2015a/b ; 2016 à par.) sont des travaux fondamentaux préparatoires sur la thématique de la présente soumission.

Pour ce qui est du TAL, l'équipe d'Osnabrück possède également une solide expérience de la constitution d'archives et de corpus. En collaboration avec l'Institut des études romanes de l'Université d'Osnabrück (C. Bürgel) et l'Université de Cologne (S. Diwersy), l'équipe a constitué le premier *corpus de référence du français contemporain* (Siepmann/Bürgel 2014).

***Bonn***

Dès les débuts de son activité d'enseignant-chercheur au sein de l'unité de recherche linguistique du département d'Études anglophones de l'Université de Cologne, M. Gymnich a accumulé une expérience significative à l'interface entre études littéraires, culturelles et linguistiques. L'intérêt qu'elle y porte se manifeste clairement dans son HDR (« Réflexions métalinguistiques et moyens d'expression dans le roman anglophone postcolonial et interculturel ») qui recourt à des catégories analytiques relevant aussi bien des études littéraires que de la linguistique et qui procède à une analyse systématique de l'emploi de différentes variétés linguistiques par différents personnages. Depuis sa thèse, publiée en 2000, le roman anglophone (anglais, américain et post-colonial) du XXe siècle constitue l'un de ses principaux champs de recherche, qu'elle envisage sous l'angle historique aussi bien que sous celui des genres, tels que le roman de formation féminin et le roman policier. Au cœur de sa réflexion sur la théorie littéraire se trouvent la théorie des genres (p.ex. 2007, 2010, 2011) et la narratologie (p.ex. 2002, 2013). En raison de son orientation comparatiste et l'attention qu'il accorde aux structures linguistiques, le projet pourra bénéficier de l'expertise du Centre de narratologie transculturelle de Bonn, dont M. Gymnich est un des membres fondateurs et membre du conseil académique. Ce centre transdisciplinaire regroupe des représentants de différentes disciplines des sciences du langage.

***Erlangen***

Les travaux menés par L. Fesenmeier portent sur la synonymie distinctive et la linguistique contrastive (français, italien, espagnol). Partant du constat que la plupart des travaux consacrés au phénomène de la synonymie ne prennent en compte que l'aspect paradigmatique (cas typique : *savoir* / *connaître*), L. Fesenmeier consacre sa thèse d'habilitation à des synonymes (p.ex. *se souvenir* / *se rappeler*) qui présentent des intersections plus ou moins considérables concernant leur comportement collocationnel (cf. Fesenmeier 2009 ; ). Dans une démarche inductive et se fondant sur de grands corpus journalistiques et littéraires, L. Fesenmeier (2008a,b, 2010) et A. Grutschus / L. Fesenmeier (2013) montrent que la différence synonymique résulte de sa conceptualisation au niveau linguistique. Les analyses menées au niveau contrastif (Fesenmeier 2008, 2010, 2013) ont montré que les

langues romanes sont très différentes au niveau des possibilités combinatoires de certains couples synonymiques (*saber* / *sapere*, *conocer* / *conoscere)*. Il est évident qu'une telle constellation lexico-grammaticale comporte des enjeux importants au niveau de la traduction. Le projet s'inscrit également dans le cadre des recherches menées au sein du *Centre Interdisciplinaire 'Dialectes et variation linguistique'* et du *Centre Interdisciplinaire Lexicographie, valence, collocation* qui se sont établis à la FAU et dont L. Fesenmeier et D. Siepmann sont membres.

***Cologne***

L'intérêt de l'équipe porte surtout sur l'élaboration de méthodologies ayant recours à une exploitation statistique de données textuelles. Dans ces travaux, une place importante est accordée à la création de corpus massifs pour différentes langues, à leur traitement automatique et au développement d'outils de requête et de calcul lexico-statistique, dont une partie a fourni un support méthodologique à d'autres projets de recherche comme Emolex (Diwersy et *al*., 2015). Une partie des activités actuelles est consacrée à la mise en place d'un vaste corpus de presse francophone (Varitext http://syrah.uni-koeln.de/varitext/). D'autres travaux, réalisés dans le cadre du projet franco-allemand PRESTO, portent sur l'élaboration d'une méthodologie visant à explorer des colligations textuelles. Enfin, l'équipe, dans une coopération étroite avec le LIDILEM a participé au développement de la plateforme multilingue EmoBase. Parmi les articles qui présentent les différents aspects des recherches mentionnées, on peut citer Diwersy (2012a, 2012b), Diwersy & Fesenmeier (2006), Diwersy & François (2011), Diwersy & Kraif (2013), Diwersy, Evert & Neumann (2014).

Les compétences des chercheurs des équipes françaises et allemandes sont donc complémentaires : en linguistique, en stylistique, en théorie des genres romanesques, en TAL. Elles sont nécessaires à la réalisation de ce projet interdisciplinaire et interlinguistique (français, anglais, allemand). Les multiples collaborations fructueuses, établies depuis plusieurs années, entre les chercheurs français et allemands, d'une part, et les nouvelles collaborations qui vont se mettre en place avec des collègues stylisticiens et des chercheurs des universités de Caen, d'Erlangen et de Bonn sont un gage du succès du présent projet. Les participants extérieurs au projet - G. Philippe (spécialiste en stylistique française), P. Blumenthal (spécialiste en linguistique contrastive), F. Maniez et J. D. Gallagher (spécialistes en traductologie), T. Muryn et M. Niziolek (spécialistes en phraséologie contrastive spécifique du roman policier) – contribueront par leur expertise à la réalisation des différents objectifs de ce projet.

## 3. OBJECTIFS ET PROGRAMME DES TRAVAUX

### 3.1. Objectifs

Nous poursuivons les objectifs suivants :

**a)** sur le plan de la recherche en TAL, extraire de façon inductive des motifs lexico-syntaxiques à partir des expressions complexes automatiquement repérées dans l'architecture implémentée dans la base de données EmoBase (cf. Figure 1) ;

**b)** sur le plan des applications pratiques, ajouter des fonctionnalités à Emoconc (module de l'EmoBase): 1) des outils permettant de mesurer la spécificité des constructions relativement aux différents genres, sous-genres et types de discours et 2) des outils permettant de rechercher des motifs séquentiels transphrastiques ;

**c)** sur le plan de la recherche linguistique, élaborer une typologie structurelle et fonctionnelle des constructions lexico-syntaxiques spécifiques (CLS) au discours romanesque francophone, anglophone et germanophone ; à partir de l'extraction de ces constructions, on spécifiera les contextes discursifs spécifiques (p.ex. rôles narratifs, changement de point de vue, séquences descriptives (cf. p.ex. Adam 2005) ;
**d)** sur le plan de la recherche stylistique : 1) asseoir une méthodologie outillée, renouvelant et étayant les approches stylistiques empiriques du texte littéraire ; 2) poursuivre la reconnaissance et la description des patrons stylistiques caractéristiques de l'écriture littéraire par comparaison entre des corpus de différents statuts (littéraires, paralittéraires, non littéraires) ; 3) approfondir la réflexion sur la poétique historique des formes en identifiant des récurrences lexico-syntaxiques caractéristiques de différents sous-genres, voire des points de convergence entre ces derniers ;
**e)** conjoindre les approches linguistiques et stylistiques décrites sous c) et d) en aboutissant à une description des patrons stylistiques littéraires en termes de constructions lexico-syntaxiques (*il en était là de* + DetPoss + GN [réflexions, pensées, discours] *quand* + proposition) et de motifs séquentiels ;
**f)** proposer des comparaisons interlinguistiques ponctuelles de CLS et d'œuvres traduites, en travaillant, d'une part, sur un corpus comparables d'œuvres françaises, anglaises et allemandes et, d'autre part, sur des corpus parallèles (de traduction) ;
**g)** jeter les fondements d'un « lexique-grammaire » des unités lexicales étendues qui contribuent à la construction de l'intrigue (scripts et trame narrative, Baroni 2002) et, plus généralement, du discours romanesque.

### 3.1.1. Aspects innovants ; retombées scientifiques et sociétales du projet

Ce projet présente un caractère innovant, à la fois par son objet, ses implications théoriques et la méthodologie qui sera mise en œuvre. Il permettra de mettre en synergie les apports des stylisticiens et des linguistes à l'étude des textes littéraires, en vue d'accéder à une connaissance plus précise de la nature linguistique des genres. Les retombées scientifiques et sociétales d'un tel projet sont multiples :
- sur le plan linguistique, il aboutira à la mise en place d'une typologie des CLS spécifiques à un type de discours, où celles-ci ont été peu étudiées (littéraire et paralittéraire) et ce, à partir de leurs propriétés linguistiques (sémantiques, syntaxiques et discursives) ;
- sur le plan du TAL, il permettra de développer des techniques nouvelles pour la fouille de données en linguistique et en stylistique et, plus particulièrement, pour l'extraction des unités polylexicales, impliquant l'identification de sous-arbres syntaxiques récurrents à partir de corpus analysés syntaxiquement ;
- sur le plan de la linguistique de corpus, il proposera une méthodologie nouvelle pour l'exploitation et l'analyse des données mono- et interlinguistiques.
- sur le plan des études stylistiques, les CLS spécifiques seront analysées avec les outils du TAL, ce qui permettra de travailler sur de grandes masses de textes, impossibles à dominer de manière non automatique. Cette nouvelle approche permettra de mieux décrire la langue littéraire dans ses formes spécifiques et dans son dialogue avec les discours non littéraires (repérer d'éventuels emprunts ou, au contraire, sa singularité radicale).
- sur le plan de la théorie des genres, il aboutira au développement d'approches permettant de mieux distinguer littérature et paralittérature et d'apporter un éclairage nouveau sur la

question de l'hybridité générique ; selon Rastier (2011 : 72), « [l]a demande sociale d'une théorie opératoire des genres est croissante ».

**-** du point de vue de la linguistique contrastive et de la traductologie, le projet permettra aussi d'asseoir la stylistique comparée et la traduction littéraire sur des bases empiriques solides ;

**-** sur le plan sociétal, il débouchera sur la mise en accès libre, via une interface Web, des données issues du projet sous forme d'une base de données interrogeable à des fins scientifiques et dans le respect des droits d'auteurs. Le projet aura également des retombées positives en ce qui concerne le transfert des résultats scientifiques vers le domaine de l'enseignement de la linguistique, de la stylistique ainsi que de la traductologie et la pratique traduisanteEn résumé, il s'agit d'un sujet original concernant un champ de recherche encore peu exploré : « Beaucoup – énormément – reste à faire dans ce domaine, puisque le chantier est encore à peine ouvert » (Legallois 2012b : 50).

### 3.2. Méthodologie, programme des travaux, calendrier

#### 3.2.1. Méthodologie de la constitution et du traitement des corpus

***La composition des corpus***

La composition de nos corpus monolingues comparables prendra comme point de départ la classification adoptée par le Brown Corpus selon des critères externes entre « belles lettres » (la littérature proprement dite) et « general fiction » (une production romanesque contemporaine moins considérée par la critique). Au fur et à mesure que le projet évoluera, cette classification sera revue et améliorée à la lumière des débats actuels autour de la distinction des genres littéraires et des analyses linguistiques selon des critères internes (cf. Lee 2001). S'y ajoute une deuxième distinction, plus fine celle-ci, entre plusieurs (sous-)genres (romans de science-fiction, policiers, sentimentaux). Nous partons de corpus comparables de même taille (environ 10 M. de mots de « littéraire » et 20 millions de mots de « paralittéraire » pour chaque langue[1]) et qui couvrent la même période (XX$^{e}$ s.). Ce décalage dans la taille des corpus s'explique par le fait que les romans d'auteur sont en quantité restreinte comparés au nombre de romans paralittéraires. Les métadonnées permettront de partitionner le corpus à différents degrés de granularité et/ou d'abstraction (depuis le texte individuel jusqu'aux sous-genres)

Pour les besoins de notre étude traductologique, on se fondera sur les corpus parallèles constitués dans le cadre d'Emolex que nous envisageons d'enrichir en respectant les critères génériques retenus pour les corpus comparables. Ces corpus comprendront environ 10 millions de mots répartis de manière équilibrée entre les couples de langue (français-anglais, français-allemand, anglais-allemand).

Les corpus journalistiques issus du projet Emolex d'environ 100 M. de mots pour chaque langue et les textes scientifiques du corpus Scientext d'environ 5 M. de mots (qui seront enrichis de textes scientifiques allemands) serviront de corpus de contraste afin de dégager les CLS spécifiques au discours romanesque (pour les différents types de corpus dont nous disposons et qui vont servir de base à la constitution des corpus de Phraseorom, cf. *Annexe* 1).

***Le traitement et l'exploitation des corpus***

Après la mise en place d'un corpus structuré (obtenu par balisage en XML et annotation (morpho-)syntaxique ; voir la section programme), on procédera à l'extraction des

[1] Les corpus français, anglais et allemands dont nous disposons dans l'Emobase seront complétés et équilibrés.

constructions lexico-syntaxiques selon une méthodologie novatrice et à l'aide d'outils informatiques sophistiqués.

Nous chercherons à développer les méthodes visant à extraire les patrons à partir desquels seront identifiées les CLS, en associant les nœuds des sous-arbres récurrents non plus à des unités lexicales, mais à des classes d'unités (classes morphosyntaxiques, mais aussi classes sémantiques – sur des bases distributionnelles ou à partir de ressources externes du type WORDNET).

En nous fondant sur des partitionnements en différents genres et sous-genres, littéraires et non littéraires, nous prévoyons d'identifier quelles sont les CLS récurrentes les plus discriminantes pour la classification des textes. Nous prévoyons de mettre en œuvre des méthodes multivariées ainsi que des calculs de spécificité fréquentielle (cf. Lebart & Salem 1994).

**3.2.2. Méthodologie de l'analyse linguistique**

Nous nous proposons, en premier lieu, de repérer les différentes CLS spécifiques au discours romanesque francophone, anglophone et germanophone, sur la base de calculs statistiques opérant à deux niveaux : d'une part, la récurrence significative des CLS à l'intérieur de l'ensemble des corpus littéraires et paralittéraires et d'autre part, les spécificités de fréquence dans les corpus littéraires, paralittéraires et non-littéraires. Après cette étape d'identification, nous établirons des classes sémantico-fonctionnelles sur la base de critères notionnels (p.ex. temps, espace, mouvement, etc.) pour l'ensemble de nos corpus et pour les différents genres. On observera par la suite s'il existe des patrons syntaxiques spécifiques à ces classes. Enfin on étudiera quel rôle jouent ces dernières dans l'organisation discursive des romans. Cette analyse s'inspire du modèle des unités lexicales étendues de Sinclair (2004) qui articule les quatre niveaux - lexical, sémantique, syntaxique et pragmatique - dans le but d'affiner ce modèle, surtout dans la perspective de l'organisation du texte littéraire. L'approche adoptée s'appuiera également sur les théories fonctionnelles (Halliday 2004, Dik 1997, Hengeveld & Mackenzie 2010). S'y ajoutent la perspective des collocations et colligations textuelles issue du modèle du *Lexical Priming* de Hoey (2005) et, dans le cas des collocations binaires, le modèle fonctionnel théorique et méthodologique développé dans Emolex.

Le deuxième volet de l'analyse linguistique est celui de l'analyse contrastive et traductologique, qui sera axée sur deux types de méthodes éprouvées. Premièrement, en partant des classes sémantico-fonctionnelles établies lors de l'analyse linguistique et qui serviront de base pour une comparaison d'ordre conceptuel, nous étudierons les relations d'équivalence entre un certain nombre de constructions de chaque classe. Ce faisant, nous testerons sur une base empirique plus solide certaines hypothèses émises par la stylistique comparée. Deuxièmement, nous nous proposons d'analyser, à l'aide des corpus de traduction, dans quelle mesure les traducteurs littéraires parviennent à rendre de façon adéquate les phraséologismes mis en perspective lors de l'analyse contrastive. Par le croisement de données fournies par les corpus monolingues et les corpus de traduction, on déterminera à quel interdiscours les traducteurs empruntent dans la langue d'arrivée.

Les comparaisons français-anglais seront prises en charge par Osnabrück et Grenoble, allemand-anglais par Osnabrück, enfin celles entre le français et l'allemand par Osnabrück et Erlangen.

### 3.2.3. Méthodologie de l'analyse stylistique

L'extraction de données permise à grande échelle par les outils de la linguistique de corpus permettra de constituer autant d'observables dont l'analyse stylistique devra déterminer la pertinence littéraire à travers un travail double de contextualisation et de comparaison. Il s'agira tout d'abord de repérer les phénomènes récurrents dans les corpus littéraires et paralittéraires. Pour analyser la pertinence proprement littéraire de ces données, il conviendra de comparer les résultats avec ce que l'on observe dans les corpus contemporains non littéraires. On sera ainsi à même de sous-catégoriser deux types d'observables : ceux qui, parce qu'ils sont surreprésentés en contexte littéraire, construisent une langue littéraire spécifique ; ceux qui n'ont pas de caractéristiques proprement littéraires. De là, les étapes de l'analyse des éléments constitutifs de la langue littéraire sont les suivantes:

- d'une part, on étudiera la valeur contextuelle de ces constructions : quels sont leur pertinence et leur rôle dans la poétique du roman ? On s'attachera notamment à deux domaines d'investigation : la « tension narrative » (Baroni 2007, Adam 2005) et la construction d'effets point de vue (Rabatel 1998, 2008). Est-il possible de leur assigner des rôles communs, par-delà la partition entre différents sous-genres ?
- d'autre part, à travers une comparaison interne, on observera s'il existe des traits caractéristiques de tel ou tel sous-genre et si, le cas échéant, la valeur de ces traits dépend de leur apparition au sein de tel ou tel sous-genre. À l'horizon même de ce travail, une réflexion sur la pertinence de ces sous-classifications se dessine.
- chemin faisant, on sera nécessairement amenés à étudier les types de corrélations qui existent entre les phraséologismes employés dans les romans de grande diffusion et ceux dont se sert la langue littéraire: y a-t-il sous-représentation ? surreprésentation ? Derrière la réponse à ces questions se perçoit un enjeu pragmatique : le roman populaire cherche-t-il à ressembler ou à s'éloigner des traits les plus caractéristiques de la langue littéraire ? Est-il linguistiquement ou thématiquement paralittéraire ?
- enfin, à partir de la comparaison entre discours littéraires et discours non littéraires, on comparera la diffusion de certaines formes : l'analyse stylistique devra alors déterminer s'il y a là influence d'un type de discours sur un autre (et dans quel sens : les discours littéraires empruntent-ils aux discours non littéraires ou inversement ?). Une périodisation du corpus sera alors nécessaire pour contribuer à décrire l'interdiscours caractéristique de tel ou tel moment.

### 3.2.4. Programme des travaux

*Programmation des tâches en TAL* :

**Tâche 1 : Corpus**

**T1.1** Collecte de nouveaux textes en sus du corpus Emolex, afin d'équilibrer les quantités au niveau des genres, pour chaque langue.

**T1.2** Reformatage en suivant le schéma XML défini pour Emolex. Regroupement des textes par blocs de fichiers XML suffisamment grands (~ 2 go) pour optimiser l'utilisation du cache dans l'architecture d'EmoConc.

**T1.3** Annotation morpho-syntaxique (étiquetage morphosyntaxique, lemmatisation et balisage des dépendances).

**T1.4** Annotation sémantique (à partir d'un thésaurus à large couverture). Ajout, pour chaque forme, d'un descripteur thématique général (issu du Thesaurus de Larousse).

**T1.5** Création des hachages enregistrant les phrases ainsi que les cooccurrences lexico-

syntaxiques, et dotés d'index par lemme, par relation de dépendance, descripteur thématique, et par méta-données (genre, auteur, année).

**Tâche 2 : Réingénierie EmoConc**

**T2.1** Dossier de conception du nouvel outil. Design des interfaces et définition des nouvelles fonctionnalités d'interrogation du corpus.

**T2.2** Modification de la structure des index, intégration de tous les niveaux de balisage structurel (paragraphe, texte, etc.).

**T2.3** Intégration des critères de partitions du corpus de travail.

**T2.4** Intégration de contraintes sur les méta-données dans le langage de requête.

**T2.5** Modification des sorties en affichant des calculs de spécificité pour chaque sous-ensemble de la partition définie sur le corpus de travail.

**T2.6** Réalisation de l'interface de définition de la partition du corpus de travail (en fonction des critères de genre, auteur, année).

**Tâche 3 : Extraction des motifs séquentiels et lexico-syntaxiques**

**T3.1** Mise en œuvre d'un algorithme d'extraction des motifs lexico-syntaxiques à partir des expressions polylexicales automatiquement extraites pour un pivot donné.

**T3.2** Création d'une interface pour la définition de motifs séquentiels extra-phrastiques.

**T3.3** Implémentation d'un script de recherche de motifs séquentiels extra-phrastiques.

***Programmation des tâches en linguistique***

**Tâche T1 :** répertorier les constructions lexico-syntaxiques extraites des corpus.

**Tâche T2** : mettre en rapport ce premier recensement avec les théories contextualistes et fonctionnelles.

**Tâche T3 :** mettre en place une grille sémantico-fonctionnelle et syntaxique pour classifier de manière fine les constructions répertoriées.

**Tâche T4 :** sélectionner un certain nombre de catégories (p.ex. temps, espace, corps, mouvement) afin d'étudier en détail leur fonctionnement dans le discours romanesque de différents genres.

**Tâche T5** : établir, pour les catégories sélectionnées dans la tâche 4, les (non-)équivalences entre les classes de constructions lexico-syntaxiques dans les trois langues comparées.

**Tâche T6** : comparer un certain nombre de traductions afin d'observer les choix des traducteurs dans le domaine des constructions lexico-syntaxiques.

**Tâche T7** : mettre en commun les résultats issus du volet linguistique et du volet stylistique.

***Programmation des tâches en stylistique et en théorie des genres***

**Tâche T1** : aider à la constitution des corpus (genres, sous-genres).

**Tâche T2** : repérer les constructions récurrentes dans les corpus littéraires et paralittéraires.

**Tâche T3** : comparer corpus littéraires et paralittéraires avec corpus non littéraires pour épingler les constructions spécifiques à la langue littéraire.

**Tâche T4** : comparer corpus littéraires et corpus paralittéraires pour identifier d'éventuelles récurrences par sous-genre et appliquer les résultats aux romans hybrides

**Tâche T5** : identifier la valeur contextuelle des constructions repérées en termes de poétique romanesque et de narratologie (tension narrative ? construction du point de vue ), discriminer différents fonctionnements caractéristiques de tel ou tel sous-corpus.

**Tâche T6 :** déterminer la spécificité des corpus paralittéraires vis-à-vis des corpus littéraires autour des usages de la langue littéraire : quel statut linguistique / quel imaginaire de la langue pour la paralittérature ?

**Tâche T7** : interpréter au sein d'une réflexion relevant de la poétique historique et de l'histoire des genres les rapports entre discours littéraires et discours non littéraires

### 3.2.5. Calendrier et répartition des travaux entre les équipes

| | 1-2 | 3-4 | 5-6 | 7-8 | 9-10 | 11-12 | 13-14 | 15-16 | 17-18 | 19-20 | 21-22 | 23-24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| T1.1 Collecte | | | | | | | | | | | | |
| T1.2 Reformatage | | | | | | | | | | | | |
| T1.3 Annotation morphosyntaxique | | | | | | | | | | | | |
| T1.4 Annotation sémantique | | | | | | | | | | | | |
| T1.5 Indexation | | | | | | | | | | | | |
| T2.1 Dossier de conception | | | | | | | | | | | | |
| T2.2 Scripts d'indexation | | | | | | | | | | | | |
| T2.3 Intégration partition corpus | | | | | | | | | | | | |
| T2.4 Ajout contraintes métadonnées | | | | | | | | | | | | |
| T2.5 Calcul de spécificité | | | | | | | | | | | | |
| T2.6 Interface pour définir la partition | | | | | | | | | | | | |
| T3.1 Motifs lexico-syntaxiques | | | | | | | | | | | | |
| T3.2 Interface pour définir les motifs séquentiels | | | | | | | | | | | | |
| T3.3 Extraction des motifs séquentiels | | | | | | | | | | | | |
| Tests et débogages | | | | | | | | | | | | |

Tableau 1 : **Planification des tâches liées au traitement des corpus**

Responsables : Olivier Kraif et Sascha Diwersy
Personnels permanents impliqués dans la réalisation de ce volet du projet:
LIDILEM : O. Kraif, ; COLOGNE : S. Diwersy.
Personnels non-permanents impliqués :
LIDILEM : 1 contrat de 12 mois + 1 contrat de 13 mois type Ingénieur de recherche.

Le temps consacré à la réalisation de ce volet : 24 mois (+ 1 mois pour la mise en fonctionnement de *PhraséoBase*).

Les livrables : Corpus multilingues annotés, scripts d'indexation et de recherche, interface d'interrogation et d'affichage des résultats, scripts d'extraction de motifs séquentiels, interface de définition de ces motifs et d'affichage des extractions ; évolution et enrichissement de l'interface de l'EmoBase afin d'y intégrer un nouveau module *PhraséoBase* (en accès libre) qui regroupera les données issues du projet .

| Mois (au total 36) | 1-9 | 10-12 | 13-15 | 16-18 | 19-24 | 24-30 | 30-36 |
|---|---|---|---|---|---|---|---|
| **T1a** Étude pilote sur corpus existants | | | | | | | |
| **T1b** Recensement des CLS sur les corpus définitifs | | | | | | | |
| **T2** Choix théoriques | | | | | | | |
| **T3** Grille sémantico-fonctionnelle et syntaxique | | | | | | | |
| **T4** Analyse discursive | | | | | | | |
| **T5** Équivalences interlinguistiques | | | | | | | |
| **T6** Comparaisons traductologiques | | | | | | | |
| **T7** Mise en commun des résultats | | | | | | | |

Tableau 2 : **Planification des tâches liées à l'analyse linguistique**

Responsables : Iva Novakova et Dirk Siepmann

Personnels permanents impliqués dans la réalisation de ce volet du projet :
LIDILEM : I. Novakova, A. Tutin, F. Grossmann, J. Sorba; D. Legallois (CRISCO)
OSNABRÜCK : D. Siepmann
ERLANGEN : L. Fesenmeier
BONN : M. Gymnich
COLOGNE : S. Diwersy

Personnels non-permanents impliqués :

LIDILEM : 1 post-doc 12 mois; 1 contrat CDD de 18 mois.
OSNABRÜCK : 1 collaborateur (Doktorand/Habilitand) 36 mois.
ERLANGEN : 1 collaborateur (Doktorand/Habilitand) 36 mois.

Le temps consacré à la réalisation de ce volet: 36 mois.

Les livrables : étude pilote sur les corpus, recensement des données des corpus comparables et parallèles, élaboration des fichiers Excel pour le traitement linguistique des données : codages syntaxico-sémantiques des résultats extraits des corpus sous forme de grilles ; rapports intermédiaires suite aux réunions rendant compte des avancées et des difficultés rencontrées ; rapport final, participations à des Colloques internationaux et publications pour la valorisation du projet ; organisation d'un Colloque international ; publication d'un ouvrage issu du projet.

| Mois | 1-16 | 16-22 | 22-24 | 24-30 | 30-36 |
|---|---|---|---|---|---|
| **T1** Constitution des corpus TAL | | | | | |
| **T2** Repérage des récurrences au sein des corpus littéraire et paralittéraire | | | | | |
| **T3** Repérage des données spécifiques aux discours littéraires par comparaison avec les discours non littéraires | | | | | |
| **T4** identification des récurrences par sous-genres | | | | | |
| **T5** interprétation de la valeur des formes récurrentes (poétique du roman) | | | | | |
| **T6** identification de l'éventuelle spécificité des sous-genres paralittéraires et interprétation des données | | | | | |
| **T7** identification et interprétation de la circulation des formes entre discours littéraires et discours non littéraires | | | | | |
| **T8** Mise en commun des résultats | | | | | |

Tableau 3 **Planification des tâches liées à l'analyse stylistique**

Responsables : Julien Piat ; Marion Gymnich
Personnels permanents impliqués dans la réalisation de ce volet du projet :
LITT & ARTS, équipe TRAVERSES 19-21 : L. Gonon et J. Piat.
BONN : Marion Gymnich

Personnels non-permanents: 1 contrat CDD de 12 mois.
BONN : 1 collaborateur (Doktorand/Habilitand) 36 mois.

Le temps consacré à la réalisation de ce volet : 36 mois.
Les livrables : participation à la constitution des corpus français et anglais ; vérifications et analyses des données extraites présentées sous forme de fichiers Excel dans la perspective stylistique, participation à des Colloques internationaux, articles à valeur de rapport final, participation à l'ouvrage issu du projet.

### 3.2.6. Diffusion des résultats

Nous envisageons une série de publications dans les revues spécialisées, des communications à des colloques nationaux et internationaux. Nous organiserons également un colloque international interdisciplinaire en Allemagne sur la phraséologie et la littérature. Ces activités de diffusion des résultats du projet nous permettront aussi d'atteindre un objectif plus ambitieux encore, à savoir la publication d'un volume international sur le thème. Ce volume inclura la globalité des résultats et des démarches issus du projet, mais aussi ceux d'autres équipes internationales avec lesquelles nous collaborons. Notre ambition est que ce volume devienne un ouvrage de référence pour les chercheurs de différentes disciplines (lexicologues, spécialistes en TAL, stylistes, traductologues)

## 3.3. BIBLIOGRAPHIE


Adam, J.-M. ([2005] 2008). *La Linguistique textuelle. Introduction à l'analyse textuelle des discours*, Paris: Armand Colin.

Ballard, M. (1992). *La traduction de l'anglais au français*. Paris : Nathan.

Baroni, R. (2002). "Le rôle des scripts dans le récit", *Poétique* n°129: 105-126.

– (2007). *La Tension narrative. Suspense, curiosité et surprise*, Paris : Editions du Seuil.

Barthes, R. (1966). "Introduction à l'analyse structurale des récits", *Communications* n° 8, Paris, Seuil.

Basseler, M., Nünning, A. & Schwanecke, C. (Eds.) (2013). *The Cultural Dynamics of Generic Change in Contemporary Fiction: Theoretical Frameworks and Model Interpretations*. Trier: WVT.

Beauvisage, T. (2001) « Exploiter les données morphosyntaxiques pour l'étude statistique des genres –Application aux roman policier, TAL No 43, TAL No 43 <http://www.revue-texto.net/Inedits/Beauvisage/index.html>.

Bourdieu, P. (1992). *Les règles de l'art : genèse et structure du champ littéraire*, Seuil.

Biber, D. (1993). "Using register-diversified corpora for general language studies." Computational Linguistics 19: 219-241.

Biber, D., Johansson, S., Leech, G., Conrad S, Finegan, E.(1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Blumenthal, P., Novakova, I. & Siepmann D. (Eds.) (2014). *Les émotions dans le discours. Emotions in Discourse*, Peter Lang: Berlin, 434 p.

Blumenthal, P. (2006). *Wortprofil im Französischen (Beihefte zur ZrP 332)*, Tübingen: Niemeyer.

Blumenthal, P., Diwersy, S. & Mielebacher, J. (2005). "Kombinatorische Wortprofile und Profilkontraste. Berechnungsverfahren und Anwendungen", *Zeitschrift für romanische Philologie* n°121: 49-83.

Brunet E. (1981). *Le Vocabulaire français de 1789 à nos jours,* Slatkine-Champion, Genève-Paris.

Burton, B. & Carter, R. (2006). "Literature and the Language of Literature". in: K. Brown, A. H. Anderson, L. Bauer, M. Berns, G. Hirst, & J. Miller (Eds.), *Encyclopedia of Language and Linguistics*, Vol. 7. Boston: Elsevier, (2nd ed., pp. 267-274).

Charest, S., Brunelle E. & Fontaine J. (2010). "Au-delà de la paire de mots : extraction de cooccurrences syntaxiques multilexémiques", *Actes de TALN 2010,* Montréal: 19-23 juillet 2010.

Chuquet, H. & Paillard M. (1987). *Approche linguistique des problèmes de traduction*. Paris : Ophrys.

Croft, W. (2001). *Radical construction grammar*. Oxford: Oxford University Press.

de Beaugrande, R. (2005). 'Corporate bridges' twixt text and language: twenty arguments against corpus research and why they're a right load of Old Codswallop. Internet: http://www.beaugrande.com/Corporate%20Bridges.htm.

Dik, S. (1997a). *The theory of functional grammar*. Part 1, *The structure of the clause*. 2nd ed., Berlin/New York: Mouton de Gruyter.

– (1997b). *The theory of functional grammar*. Part 2, *Complex and derived constructions,* 2nd ed., Berlin/New York: Mouton de Gruyter.

Diwersy, S. (2012a). "Kookkurrenz, Kontrast, Profil. Korpusinduzierte Studien zur lexikalisch-syntaktischen Kombinatorik französischer Substantive (mit ergänzenden Betrachtungen zum Deutschen)", *Beihefte zur Zeitschrift für romanische Philologie:* 373.

– (2012b). "La francophonie multivariée. Ou: comment mesurer les français en Afrique?", *Le français en Afrique* n°27: 75-91. [URL: http://www.unice.fr/ILF-CNRS/ofcaf/27/DIWERSY.pdf]

Diwersy, S., Evert, S. & Neumann, S. (2014). A semi-supervised multivariate approach to the study of language variation, in: B. Szmrecsanyi, & B. Waelchli, (Eds.), *Linguistic variation in text and speech, within and across languages*, Berlin e.a.: de Gruyter, 174-204.

Diwersy, S. & Fesenmeier, L. (2006). A (Twofold) Contrastive Analysis of Collocations: Fr. patience – It. pazienza, in: E. Corino, C. Marello, & C. Onesti (Eds.), *Proceedings of the 12th EURALEX International Congress*, Alessandria: Edizione dell'Orso: 959-966.

Diwersy, S. & François, J. (2011). "La combinatoire des noms d'affect et des verbes supports de causation en français: étude de leur attirance au niveau des unités et de leurs classes syntactico-sémantiques", *Tranel* n°55: 139-161.

Diwersy S., Goossens V., Grutschus A., Kern B., Kraif O., Melnikova, E. Novakova I. (2014) Traitement des lexies d'émotion dans les corpus et les applications d'*EmoBase ,* revue Corpus No 13, 269-293.

Diwersy, S. & Kraif, O. (2013). Observations statistiques de cooccurrents (lexico-) syntaxiques pour la catégorisation sémantique d'un champ lexical, in: Baider, F. & Cislaru, G. (Eds.), *Cartographie des émotions*, Paris: Presses universitaires de la Sorbonne, 55-70.

Duff, D. (2000). *Modern Genre Theory*. London: Longman.

Evert, S. (2008). Corpora and collocations. in: A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*, Berlin: Mouton de Gruyter: 1212-1249.

Feilke, H. (1994). *Common sense-Kompetenz. Überlegungen zu einer Theorie des 'sympathischen' und 'natürlichen' Meinens und Verstehens*. Frankfurt/Main: Suhrkamp.

– (1996). *Sprache als soziale Gestalt. Ausdruck, Prägung und die Ordnung der sprachlichen Typik*. Frankfurt/Main: Suhrkamp.

– (2003). „Textroutine, Textsemantik und sprachliches Wissen." In: Linke, A., Ortner, H. & Portmann P. R. (Eds.): *Sprache und mehr. Ansichten einer Linguistik der sprachlichen Praxis*. (Reihe Germanistische Linguistik 245). Tübingen (Niemeyer): 209-229.

Fesenmeier, L. (en préparation, in Vorb.). *Zentraldifferenzen im Wortschatz des Französischen, Italienischen und Spanischen* (vorauss.: Frankfurt a.M.: Klostermann; zugleich Habilitationsschrift, Köln, 2009).

– (2008a). Komplementsätze bei span. *conocer* (und *saber*), in: E. Stark, R. Schmidt-Riese & E. Stoll, *Romanische Syntax im Wandel*, Tübingen: Narr, 399-415.

– (2008b). Quasi-sinonimia e 'differenze centrali': la coppia *aspettare – attendere*, in: E. Cresti, *Prospettive nello studio del lessico italiano. Atti del IX Congresso SILFI (Firenze, 14-17 giugno 2006)*, Firenze: Firenze University Press, 267-274.
– (2009). Synonymie: Zur Semiotik der Absenz, in: A. Grutschus & P. Krilles, *Figuren der Absenz*, Berlin: Frank & Timme: 219-233.
– (2010). 'Se souvenir' en français et en italien: différence(s) de centrage, in: M. Iliescu, H. Siller-Runggaldier & P. Danler, *Actes du XXV^e Congrès International de Linguistique et de Philologie Romanes*, Tübingen: Niemeyer, Vol. 3 : 85-96.
Fillmore, C., Kay, P. & O'Connor, C. (1988). "Regularity and idiomaticity in grammatical constructions: the case of *let alone*", *Language* 64: 501-38.
Firth, J.R. (1957). "A synopsis of linguistic theory 1930-1955", in: J.R. Firth e.a., *Studies in Linguistic Analysis*, Oxford: Philological Society: 1-32.
Fischer-Starcke, B. (2010). *Corpus linguistics in literary analysis: Jane Austen and her contemporaries*. London: Continuum.
Fletcher, W.H. (2007). *kfNgram* http://www.kwicfinder.com/kfNgram/kfNgramHelp.html.
François, J. (1989). *Changement, Causation, Action – Trois catégories sémantiques fondamentales du lexique verbal français et allemand*. Genève: Droz.
Frow, J. (2006). *Genre : The New Critical Idiom*. London: Taylor and Francis.
Gallagher, J. D. (2007). Traduction littéraire et études sur corpus. in: M. Ballard & C. Pineira-Tresmontant (Eds.), *Les corpus en linguistique et en traductologie*. Arras: Artois Presses Université: 199-230.
Goldberg A. (1995). Constructions. A construction grammar approach to argument constructions. Chicago: The University of Chicago Press.
Gonon, L. (2010). "Mythes et démystification dans le roman policier de Fred Vargas", in: *Recherches et Travaux*, n°77, *Le Devenir-roman des Mythologies de Roland Barthes*, G. Bellon & P. Vachaud (dir.), Grenoble: ELLUG, 119-135.
– (2012), *Le Fait divers criminel dans la presse quotidienne française du XIX^e siècle*, Paris: Presses Sorbonne Nouvelle.
Granger S. & Lerot, J. & Petch-Tyson, S. (2003). *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, Rodopi: Amsterdam and Atlanta.
Greaves, C. (2005). *ConcGram concordancer with ConcGram analysis*, HongKong: HKUST
Greimas, A. J. (dir.) (1982 [1972]), *Essais de sémiotique poétique*, Paris: Larousse.
Grossmann, F. & Tutin, A. (2002). "Collocations régulières et irrégulières: esquisse de typologie du phénomène collocatif", *Revue française de linguistique appliquée* n°VII-1: 7-25.
Grünbeck, B. (1976). *Moderne deutsch-französische Stilistik auf der Basis des Übersetzungsvergleichs (Sammlung romanischer Elementar- und Handbücher)*, Heidelberg: Winter.
Grutschus, A. & Fesenmeier, L. (2013). 'Inter *metum*, *timorem* et *pavorem* interest ...' – et qu'en est-il des différences entre leurs successeurs romans?, in: E. Casanova Herrero & C. Calvo Rigual, *Actes del 26é Congrés de Lingüística i Filologia Romàniques (València, 6-11 de setembre de 2010)*, Berlin: de Gruyter, Vol. 3, 171-182.
Guiraud, P. (1954). *Les caractères statistiques du vocabulaire*, Paris: P.U.F.

Gymnich, M. (2000). *Entwürfe weiblicher Identität im englischen Frauenroman des 20. Jahrhunderts*. Trier: WVT, 2000.

- (2002), „Linguistics and Narratology: The Relevance of Linguistic Criteria to Postcolonial Narratology.“ In: Marion Gymnich, Ansgar Nünning & Vera Nünning (Hg.). *Literature and Linguistics: Approaches, Models and Applications*. Trier: WVT, 61-76.

- (2004a), „Metasprachliche Reflexionen und Sprachkritik im postmodernen amerikanischen Roman.“ In: Reinhard Kacianka & Peter V. Zima (Hg.). *Krise und Kritik der Sprache: Literatur zwischen Spätmoderne und Postmoderne*. Tübingen/Basel: Francke, 233-249.

- (2004b), „Konzepte literarischer Figuren und Figurencharakterisierung.“ In: Vera Nünning & Ansgar Nünning (Hg.). *Erzähltextanalyse und Gender Studies*. Stuttgart/Weimar: Metzler, 122-142.

- (2007), *Metasprachliche Reflexionen und sprachliche Gestaltungsmittel im englischsprachigen postkolonialen und interkulturellen Roman*. Trier: WVT, 2007.

- (2007), „Vorschläge für eine Relationierung verschiedener Aspekte und Dimensionen des Gattungskonzepts: Der Kompaktbegriff Gattung.“ In: Marion Gymnich, Birgit Neumann & Ansgar Nünning (Hg.). *Gattungstheorie und Gattungsgeschichte*. Trier: WVT, 31-52. (mit Birgit Neumann)

- (2010), „Gattung und Gattungshistoriographie.“ In: Rüdiger Zymner (Hg.). *Handbuch Gattungstheorie*. Stuttgart/Weimar: Metzler, 131-158.

- (2011), „‚Decolonizing Genre?‘ – Das Konzept der literarischen Gattung in englischsprachiger postkolonialer und interkultureller Literatur.“ In: Stephan Conermann & Amr ElHawary (Hg.). *Was sind Genres? Nicht-abendländische Kategorisierungen von Gattungen*. Berlin: EB-Verlag, 299-315.

- (2013), "Gender and Narratology." In: *Literature Compass* 10.9: 705-715.

Gymnich, M., Neumann, B. & Nünning, A. (2007). *Gattungstheorie und Gattungsgeschichte*. Trier: WVT.

Habert, B. (2005). I*nstruments et ressources électroniques pour le français.* Gap/Paris : Ophrys, 2005, 170 p.

Halliday, M.AK. (2004). *An Introduction to Functional Grammar*, 3th ed. London: Edward Arnold.

Hausmann, F.J. (1979). “Un dictionnaire des collocations est-il possible?”, *Travaux de littérature et de linguistique de l’université de Strasbourg n°XVII/1*: 187-195.

– (2007). *Collocations, phraséologie, lexicographie. Études 1977-2007 et Bibliographie*. Aachen: Shaker 2007.

Hengeveld, K. & Mackenzie, J. L. (2010). Functional Discourse Grammar, in: B. Heine & H. Narrog (Eds), *Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press. 367-400. Reviewed by F. J. Newmeyer in *Studies in Language* 36.1 (2012): 215-224.

Herbst, T. & M. Klotz (2003). Lexikographie: Eine Einführung. Paderborn: Schöningh.

Hoey, M. (2005). *Lexical priming: a new theory of words and language*, London/New York: Routledge.

Hunston, S. & G. Francis (2000). *Pattern Grammar*. Amsterdam: Benjamins.

Jakobson, R. (1960). “Closing Statement: Linguistics and Poetics.” In: T.A. Sebeok, Style in Language, Cambridge, MS: MIT Press: 350-377.

Kern, B. & Grutschus, A. (2014a): „Decepción, surprise, colère et furia : exploration d’une méthode statistique en lexicologie“, in: *Zeitschrift für romanische Philologie* 130/3.

– (2014b). „Surprise vs. étonnement : comportement discursif et perspectives contrastives“, in: Blumenthal P., Novakova I., Siepmann D. (Eds.), *Les émotions dans le discours. Emotions in discourse.* Peter Lang, Berlin : 187-198.

Kraif O. (2008) Comment allier la puissance du TAL et la simplicité d'utilisation ? l'exemple du concordancier bilingue ConcQuest, *JADT 2008 : actes des 9es Journées internationales d’Analyse statistique des Données Textuelles,* Presses universitaires de Lyon, vol. 2, pp. 625-634

Kraif, O. & Diwersy, S. (2012). “Le Lexicoscope: un outil pour l'étude de profils combinatoires et l'extraction de constructions lexico-syntaxiques”, *Actes de la conférence TALN 2012*, Grenoble: 399-406.

Kraif, O. & Diwersy, S. (2014). “Exploring Combinatorial Profiles Using a Parsed Corpus: a Case Study in the Lexical Field of Emotions”, in: Blumenthal P., Novakova I., Siepmann D. (Eds), *Les Emotions dans le discours. Emotions in discourse*. Peter Lang, Berlin : 381-394.

Lebart, L. & Salem, A. (1988). *Analyse statistique des données textuelles. Questions ouvertes et lexicométrie*, Paris: Dunod.

– (1994). *Statistique textuelle*. Paris: Dunod.

Lee, D. (2001) “Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle”, *Language learning and technology* 5/3: 37-72.

Leech, G. & Short, M. (2007). *Style in Fiction. A Linguistic Introduction to English Fictional Prose*. 2nd ed. London: Pearson.

Legallois, D. (2006). “Quand le texte signale sa structure la fonction textuelle d'une certaine catégorie nominale”*, Corela* n° spécial *corela.edel.univ-poitiers.fr.*

– (2012a). “La colligation : autre nom de la collocation grammaticale ou autre logique de la relation mutuelle entre syntaxe et sémantique?”, *Corpus n°11*, URL: http://corpus.revues.org/2202.

– (2012b - avec Cellier, P., Quiniou, S. & Charnois, Th.) “Fouille de données pour la stylistique: l’exemple des motifs émergents”, *Journées Internationales d’analyse statistique des données textuelles*, Liège, 13-15 juin.

Legallois D. & Tutin A. (2013). “Présentation: Vers une extension du domaine de la phraséologie”, *Langages n°1*89: 3-25.

Lits, M. (2011). *Le Roman policier dans tous ses états: d’Arsène Lupin à Navarro*, Limoges: Pulim.

Longrée, D. & Mellet, S. (2013). “Le motif : une unité phraséologique englobante? Étendre le champ de la phraséologie de la langue au discours”, *Langages n°*189 : 68-80.

Magri-Mourgues V. (2006). “Stylistique générique et statistique”. *Les Cahiers de la MSH Ledoux* 8 : 655-666.

Mahlberg, M. (2013). *Corpus stylistics and Dickens’s fiction*. Routledge.

Maingueneau, D. & Philippe, G. (2002). "Les conditions d'exercice du discours littéraire", in: E. Roulet & M. Burger (dir.), *Les Modèles du discours au défi d'un "dialogue romanesque": l'incipit du roman de R. Pinget* "Le Libera", Nancy: PUN : 351-377.

Malblanc, A. (1961). *Stylistique comparée du français et de l'allemand. Essai de représentation linguistique comparée et étude de traduction*. Paris : Didier.

Marion, F. (2009). "Le stéréotype dans le roman policier", *Cahiers de narratologie* n°17: http://narratologie.revues.org/1095

Mel'čuk, I., Clas A., & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve: Duculot (Champs linguistiques: Manuels).

Melnikova E., Novakova, I. & Kraif O. (2009). "Quels corpus pour l'analyse contrastive? L'exemple des constructions verbo-nominales de sentiment en français et en russe", *Actes des 6èmes Journées de la Linguistique de Corpus (http://web.univ-ubs.fr/corpus/jlc6.html#publi2009)*.

Monte, M. & Philippe, G. (dir.) (2014). *Genres et textes. Déterminations, évolutions, confrontations*, Lyon: PUL.

Novakova I. (2015). "Les émotions entre lexique et discours", in :Rabatel A., Ferara-Léturgie A, Létrugie A. (eds) *La sémantique et ses interfaces*, Actes du Colloque 2013 de l'ASL, Lambert-Lucas 181-204.

Novakova I. & E. Melnikova (2013). Vers un modèle fonctionnel pour l'analyse du lexique des émotions dans cinq langues européenne, in: *le Bulletin de la Société de linguistique de Paris*, Vol. CVIII (2013), fasc. 1: 131-160.

Novakova I. & Sorba J. (2013a). Argumentation et émotion dans les séquences textuelles journalistiques. Le cas de *stupeur* et de *jalousie,* in: T. Muryn, S. Mejri, W. Prazuch & I. Sfar (Eds.), *Phraséologie et littérature*, Bern: Peter Lang, 137-149.

– (2013b). *Stupéfier* et *jalouser* dans les séquences textuelles journalistiques: quel profil discursif pour quelle stratégie argumentative?, *Le discours et la langue* Vol. 4-1, (2012 [2013]): 203-220.

Novakova I. & Sorba J. (2014a): "L'émotion dans le discours. A la recherche du profil discursif de stupeur et de jalousie", in: P. Blumenthal, I. Novakova & D. Siepmann (Eds.), *Les Emotions dans le discours. Emotions in Discourse*, Berlin: Peter Lang : 161-175.

Novakova I. & Sorba J. (2014b) L'évaluation à travers les émotions : le cas d'*estime* et de *déception*, *Langues française n*°184, 74-89.

Philippe, G. (2002). *Sujet, verbe, complément. Le moment grammatical de la littérature française 1890-1940*. Paris: Gallimard.

– (2004). *Flaubert savait-il écrire? Une querelle grammaticale (1919-1921)*, Grenoble: Ellug.

Philippe, G. & Piat J. (2009). *La langue littéraire. Une histoire de la prose en France de Gustave Flaubert à Claude Simon*. Paris: Fayard.

Piat, J. (2006). "Vers une stylistique des imaginaires langagiers", *Corpus* n°5: 113-141.

– (2011), *L'Expérimentation syntaxique dans l'écriture du Nouveau Roman (Beckett, Pinget, Simon). Contribution à une histoire de la langue littéraire dans les années 1950*, Paris: Honoré Champion.

– (2013). “Langue littéraire / style : d’un *continuum* et d’une complémentarité”, in: C. Badiou-Monferran (dir.), *La Littérarité des belles-lettres. Un défi pour les sciences du texte?*, Paris: Garnier : 359-369.

Quiniou, S., Cellier, P., Charnois, T. & Legallois, D. (2012). Fouille de données pour la stylistique: l’exemple des motifs emergents, *Actes des 11es Journées Internationales d’analyse statistique des données textuelles*, Liège, 13-15 juin 2012 : 821-833.

Rabatel, A. (1998). *La Construction textuelle du point de vue*, Lausanne-Paris, Delachaux & Niestlé, «Sciences des discours».

– (2008). *Homo narrans. Pour une analyse énonciative et interactionnelle du récit*, Vol. II, *Dialogisme et polyphonie dans le récit*, Limoges : Lambert-Lucas.

Rastier F. (2011). *La mesure et le grain. Sémantique de corpus.* Honoré Champion.

Renouf, A. & Sinclair, J. (1991). “Collocational frameworks in English”, in: K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguisitcs. Studies in honour of Jan Svartvik*. London: Longman, 128-144.

Seretan V., Nerima L. & Wehrli E. (2003). “Extraction of Multi-Word Collocations Using Syntactic Bigram Compositio”, *Proceedings of the Fourth International Conference on Recent Advances in NLP*, (RANLP-2003): 424–431.

Scholes R. (1982). *Semiotics and interpretation*. New Haven: Yale University Press.

Siepmann, D. (2002). "Eigenschaften und Formen lexikalischer Kollokationen: Wider ein zu enges Verständnis", *Zeitschrift für französische Sprache und Literatur* n°112: 240-263.

– (2004). "Kollokationen und Fremdsprachenlernen: Imitation und Kreation, Figur und Hintergrund", *Praxis Fremdsprachenunterricht* 2/2004: 107-113.

– (2005a). *Discourse Markers across Languages. A Contrastive Study of Second-Level Discourse Markers in Native and Non-native Text with Implications for General and Pedagogic Lexicography.* Abingdon/New York: Routledge (Routledge Advances in Corpus Linguistics 6).

– (2005b). "Collocation, Colligation and Encoding Dictionaries. Part I: Lexicological Aspects", International Journal of Lexicography n°18, 4: 409-444.

– (2011). “Sinclair revisited: beyond idiom and open choice.” In: Herbst, Thomas, Faulhaber, Susen & Uhrig, Peter (Eds.), *The phraseological view of language: a tribute to John Sinclair*, Berlin: Mouton de Gruyter 2011, 59-86.

– (2012). *Wissenschaftliche Texte auf Englisch schreiben*. Stuttgart: Klett.

– (2014b). “Extent of collocational difference between English, French and German emotion nouns: a corpus-based study”, in: P. Blumenthal, Peter, I. Novakova & D. Siepmann (Eds.), *Les émotions dans le discours/Emotions in discourse*, 39-53.

– (2014c). “Collocations across Languages: Unity in Diversity?” In: Dominguez, M.J., Mollica, F. & M. N. Curcio (Eds.): Zweisprachige Lexikographie im Spannungsfeld zwischen Translation und Didaktik. (= Lexicographica: Series Maior 145) Berlin: New York: de Gruyter.

- (2015a) Über Vorbilder und Zerrbilder: Literarische Texte als Grundlage von Spracharbeit. W. Hallet, C. Surkamp, U. Krämer (Eds.), Literaturkompetenzen Englisch: Modellierung - Curriculum – Unterrichtsbeispiele.

- (2015b) Lexicologie et phraséologie du roman: quelques pistes pour le français et l'anglais. Soumis à *Cahiers de lexicologie*.
- (2015c) The *corpus de référence du français contemporain* as the first balanced mega-corpus of French. Soumis à *International Journal of Corpus Linguistics*.
- – (2016). "A corpus-based investigation into key words and key patterns in post-war fiction." *A paraître* dans *Functions of Language* 2016/2.

Siepmann, D. & Bürgel, C. (2014). Le corpus de référence du français contemporain. Présentation au congrès des francoromanistes, Münster, Allemagne, 29.09.14, [URL : http://zenodo.org/record/12353#.VORzHi7UJsU.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
– (2004). *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Stefanowitsch, A. & Gries S. T. (2003). Collostructions: Investigating the interaction between words and constructions, *International Journal of Corpus Linguistics* 8.2: 209-43.

Steyer, K. (2013). *Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpusanalytischer Sicht*. Tübingen: Narr.

Stubbs, M. (2005). "Conrad in the computer: examples of quantitative stylistic methods. *Language and Literature*." 14 (1): 5-24.

Stubbs, M. & Barth I. (2003). "Using recurrent phrases as text-type discriminators. A quantitative method and some findings", *Functions of Language* 10 (1): 61-104.

Talmy, L. (2000). *Towards a cognitive semantics*. Cambridge, MS: MIT Press.

Todorov, Tz. (1980). *Poétique de la prose: choix, suivi de Nouvelles recherches sur le récit*, Paris: Editions du Seuil.

Truffaut, L. (1983), *Problèmes linguistiques de traduction allemand-français. Guide de l'étudiant et du praticien*. Munich : Hueber.

Tutin, A., I. Novakova, F. Grossmann & C. Cavalla (2006). "Esquisse de typologie des noms d'affect à partir de leurs propriétés combinatoires". *Langue française n*°150: 32-49.

Tutin, A. (2010). *Sens et combinatoire lexicale: de la langue au discours*. Habilitation thesis, University of Grenoble. http://w3.u-grenoble3.fr/lidilem/labo/file/HDR_Tutin.pdf.
– (2013). Les collocations lexicales: une relation essentiellement binaire définie par la relation prédicat-argument. *Langages*, 1(189), 47-63.

Tutin, A. & Grossmann, F. (2014). *L'écrit scientifique, Du lexique au discours*, Rennes: PUR.

Vaillant, A. (2003). "Pour une histoire de la communication littéraire", *Revue d'histoire littéraire de la France* 103e année, n° 3: 549-562.

Vaudrey-Luigi, S. (2013), *La langue romanesque de Marguerite Duras. Une liberté souvenant*, Paris: Garnier.

Vinay, J.P. & J. Darbelnet (1958). *Stylistique comparée du français et de l'anglais*. *Méthode de traduction*. Paris : Didier.

Zymner, R. (2003). *Gattungstheorie : Probleme und Positionen der Literaturwissenschaft*. Paderborn: Mentis.

## 4. MOYENS DEMANDÉS

4.1. Partie allemande :

| | |
|---|---|
| a) Wissenschaftliche Mitarbeiter | |
| Universität Osnabrück: eine halbe Stelle TV-L 13/Stufe 2: | 98100 € |
| Universität Erlangen: eine halbe Stelle TV-L 13/Stufe 2: | 98100 € |
| Universität Bonn: eine halbe Stelle TV-L 13/Stufe 2 | 98100 € |
| b) Wissenschaftliche Hilfskräfte: Universität Osnabrück: | |
| eine studentische Hilfskraft/10 Std./Monat | 9 000 € |
| Universität Erlangen: eine studentische Hilfskraft/10 Std./Monat | 9 000 € |
| Universität Köln: eine studentische Hilfskraft/10 Std./Monat | 9 000 € |
| Universität Bonn: eine studentische Hilfskraft/10 Std./Monat | 9 000€ |
| c) Universität zu Köln: Werkverträge für Softwareentwicklung | 10 000 € |
| | |
| Summe Personalkosten: | 331 300 € |

L'agrandissement de l'équipe allemande par l'intégration de M. Gymnich répond à la critique formulée dans l'évaluation de notre première soumission et vise à combler le manque d'expertise littéraire soulevé dans les rapports d'évaluation et à assurer une meilleure complémentarité des équipes française et allemande sur le plan de l'analyse littéraire et stylistique. La complexité du projet impose d'ailleurs la nécessité de créer trois postes post-doc à Osnabrück, à Erlangen et à Bonn, dont les tenants se pencheront sur la phraséologie du roman, la comparaison interlinguistique et l'interface entre stylistique et théorie des genres. Les collaborateurs étudiants situés à Osnabrück, à Bonn et à Cologne aideront à la constitution et à l'annotation des corpus, à l'entraînement des logiciels, à la recherche de littérature et à la préparation éditoriale de publications. Le collaborateur étudiant situé à Erlangen contribuera à la préparation de la conférence internationale (voir 3.2.6.).

4.1.1.2 Sachmittel

4.1.1.2.1 Mittel für Geräte

| | |
|---|---|
| Universität Osnabrück: Datenserver | 2000 € |
| Universität Osnabrück: Rechner für Mitarbeiter | 1000 € |
| Universität Erlangen: Datenserver | 2000 € |
| Universität Erlangen: Rechner für Mitarbeiter | 1000 € |
| Universität Bonn: Rechner für Mitarbeiter | 1000€ |
| Summe: | 7000 € |

4.1.1.2.2 Mittel für Reisen

| | |
|---|---|
| Universität Osnabrück 3 x Reisen nach Grenoble (2 Personen) à 800 €/Reise | 4.800 |
| 2 x Teilnahme an Tagung Ausland (2 Personen) à 1100 €/Reise | 4.400 |
| 1 x Teilnahme an Tagung Inland (2 Personen) à 500 €/Reise | 1.000 |
| Universität Erlangen 3 x Reisen nach Grenoble (2 Personen) à 800 €/Reise | 4.800 |
| 2 x Reisen nach Osnabrück (2 Personen) à 500 €/Reise | 2.000 |
| 2 x Teilnahme an Tagung Ausland (2 Personen) à 1100 €/Reise | 4.400 |
| 1 x Teilnahme an Tagung Inland (2 Personen) à 500 €/Reise | 1.000 |

| | |
|---|---|
| Universität Bonn 3 x Reisen nach Grenoble (2 Personen) à 800 €/Reise | 4.800 |
| 2 x Reisen nach Osnabrück (2 Personen) à 500 €/Reise | 2.000 |
| 2 x Teilnahme an Tagung Ausland (2 Personen) à 1100 €/Reise | 4.400 |
| 1 x Teilnahme an Tagung Inland (2 Personen) à 500 €/Reise | 1.000 |
| | |
| Summe | 34 600 € |

4.1.1.2.3 Mittel für Gäste

| | |
|---|---|
| Universität Erlangen: 3 Gastredner bei internationaler Tagung à 800€ | 2 400 |

4.1.1.2.4 Publikationskosten

| | |
|---|---|
| Universität Osnabrück/Erlangen, Publikation eines Tagungsbandes | 2 250 |

4.1.1.2.5 Sonstige Kosten

| | |
|---|---|
| Vorbereitung einer internationalen Tagung | 15 000 |

**4.1.2. Modul Eigene Stelle**

**-**

**4.1.3. Modul Vertretung**

**-**

**4.1.4. Modul Mercator-Fellow**

**-**

**4.1.5. Modul Projektspezifische Workshops**

**-**

**4.1.6. Modul Öffentlichkeitsarbeit**

-

4.2. Partie française :

**4.2.1. Dépenses de personnels à recruter**

Les dépenses prévues en personnel à recruter sont réparties comme suit :

| | |
|---|---|
| - Volet TAL : Recrutement de deux ingénieurs de recherches (IR) : 1 contrats CDD x 12 mois + 1 contrat CDD 13 mois= 25 mois, | 67 700 |
| - Volet linguistique : Recrutement d'un post-doc 12 mois | 40 000 |
| - et d'un IR de 18 mois, | 49 000 |
| - Volet stylistique : recrutement d'un IR de 12 mois | 32 700 |
| Total : | **189 400** |

Les compétences requises pour le recrutement des 2 IR en TAL : lemmatisation, annotation morpho-syntaxique des corpus, connaissances dans le domaine des applications logicielles appropriées pour la constitution, le traitement des corpus et la mise en place d'une base de données (cf. tableau 1, tâches T1.1. à T1.5. ; T3.1. et T3.3.).

Les compétences requises pour le Post-doc en linguistique : spécialiste en linguistique de corpus et en phraséologie qui participera à l'étude pilote sur les corpus (T1) , à l'analyse linguistique des données recensées sur les plans syntaxique, sémantique et discursif (T3 et T4). La connaissance de l'anglais est une condition souhaitable au recrutement dans l'objectif de la T5 (équivalences interlinguistiques français-anglais). Il participera à la coordination des tâches du volet linguistique et du volet stylistique (T7).

Les compétences requises pour l'IR en linguistique (CDD de 18 mois ) : recensement des données des corpus comparables et parallèles (français-anglais) (T1); élaboration des fichiers Excel pour le traitement linguistique des données : codages syntaxico-sémantiques des résultats extraits des corpus (T1 et T3), ainsi que participation à la T7.

Les compétences requises pour l'IR en stylistique (CDD de 12 mois) : participation aux repérages des données spécifiques aux corpus littéraires en contraste avec le corpus journalistique (T2 et T3) ; participation à la mise en commun des résultats (T8).

Les différences dans les montants prévus pour les dépenses de personnels à recruter dans le budget français et allemand sont liées aux différentes réglementations de l'ANR (le taux des CDD ne devant pas être supérieur à 30%) et de la DFG pour ce poste du budget.

**4.2.2. Équipements**

| | |
|---|---|
| Achat d'ordinateurs portables (10 x 1 000 euros) | 10 000 |
| 1 serveur site Internet | 2 000 |
| Total : | **12 000** |

**4.2.3. Petits matériels, consommables, fonctionnement, etc.**

Les frais de fonctionnement (charges externes + facturation interne) demandés seront alloués à l'achat du matériel nécessaire à la constitution des corpus et au traitement des données, ainsi qu'à l'achat d'ouvrages.

| | |
|---|---|
| Achat de disques durs externes | 2 500 |
| Achat de logiciels | 4 000 |
| Consommables, bureautique | 4 000 |
| Achat de livres, d'ouvrages spécialisés | 3 500 |
| Photocopies, numérisations de documents | 3 500 |
| Total | 17 500 |
| Stagiaires 6 ( 2 TAL, 2 volet linguistique, 2 pour la stylistique) x 6 mois x 555 euros | 20 000 |

La somme prévue pour la gratification des stagiaires (niveau Master 2) est éligible dans le poste Fonctionnement du budget (selon les règles de l'établissement). Les stagiaires vont participer et apporter leur aide à l'accomplissement des tâches multiples dans les volets respectifs du projet tout en se formant à la recherche scientifique.

Total : Fonctionnement (Autres dépenses) **37 500**

**4.2.4. Missions**

Le coût des missions s'élève à 19% de la demande financière. Pendant les trois années du projet, les membres des équipes françaises et les coordinateurs se réuniront à des réunions générales. Nous prévoyons deux réunions générales par an (1 en France en mars, 1 en

Allemagne en octobre). Les thèmes prévisionnels des réunions pourront être précisés comme suit :

**Réunion 1** : Enrichissement des corpus, perfectionnement des outils d'interrogation, répartition des tâches (informatiques, linguistiques et stylistiques)

**Réunion 2** (Allemagne): Coordination des méthodologies (mise en place des grilles d'analyses linguistiques et stylistiques pour les trois langues). Mise en place d'une étude pilote linguistique et stylistique sur des échantillons choisis des corpus

**Réunion 3** (France): Point étape sur le traitement, l'extraction, le codage et l'analyse des données

**Réunion 4** (Allemagne): Point étape sur le traitement, l'extraction, le codage et l'analyse des données

**Réunion 5** (France): Mise en commun des résultats des codages et des analyses linguistiques et stylistiques. Elaboration de la typologie des CSL spécifiques

**Réunion 6** (Allemagne) : Mise en contraste des CLS dans les trois langues comparée. Mise en place de la base de données. Modalités d'organisation du Colloque international et de la publication.

Le coût est estimé à 800 € en moyenne par personne et par mission (400 € hébergement et repas + 400 € de transport). Il est aussi prévu d'organiser 3 réunions partielles des coordinateurs et de certains membres de l'équipe en fonction du calendrier des tâches et de leur état d'avancement.

La dimension internationale et multilingue du projet implique aussi la participation active des membres des équipes à des Colloques internationaux afin de présenter et diffuser les résultats du projet. Une justification importante des moyens demandés est d'offrir à tous les membres du projet la possibilité d'internationaliser leurs travaux et de diffuser leurs résultats. L'expérience prouve que le coût moyen du séjour dans un congrès international à l'étranger est de 1100 euros environ par personne (transport, hébergement, frais d'inscription). Nous prévoyons 2 Colloques internationaux et 1 Colloque en France par personne pour toute la durée du projet. Le coût des colloques en France est estimé approximativement à 700 euros en moyenne par participant (250 € déplacement, 300€ hébergement et repas, 150 € frais d'inscription). Afin de pouvoir mener à bien le projet, il est également nécessaire de prévoir au moins 4 invitations des collaborateurs externes (de Pologne, Suisse, France, Allemagne, cf. liste des participants externes). En résumé :

| | |
|---|---|
| 3 réunions générales pour 8 personnes en Allemagne x 800 € | 19200 |
| 3 réunions partielles pour 2 personnes x 800 € | 4800 |
| 2 Colloques internationaux par personne 8 x 1100 € | 17600 |
| 1 Colloque national par personne 8 x 700 € | 5600 |
| Participation au Colloque final en Allemagne : 8 x 1100 € | 8800 |
| Invitations de collaborateurs externes au projet | 4000 |
| Déplacements de D. Legallois (Caen-Grenoble) 3 x 600€ | 1800 |
| Total | **61 800** |

**NB**. Les missions et les dépenses de matériels pour D. Legallois (Crisco, U. Caen) font partie intégrante du budget du Lidilem (porteur du projet), car n'excédant pas 15 000 euros sur trois ans.

Les sommes indiquées dans les rubriques de 4.2.1 à 4.2.4. globalisent le budget prévu pour les deux partenaires français (Lidilem et Litt&Arts équipeTraverses 19-21).
La différence dans les montants globaux des missions dans le budget français et allemand s'explique par le plus grand nombre des chercheurs permanents impliqués dans les équipes françaises.

**4.2.5. Prestations de services**

| | |
|---|---|
| Sur facturation externe : | |
| Développement et mises à jour d'un site Web pour le projet | 3000 |
| Traduction d'articles, avance sur publication d'ouvrage, autres | 9 000 |
| Total | **12 000** |

**5. RECAPITULATIF DE LA DEMANDE FINANCIERE**

| | FR | | D |
|---|---|---|---|
| Dépenses de personnel | 189 400 | Basismodul: Personalkosten<br>Modul Eigene Stelle<br>Modul Vertretung<br>Modul Mercator-Fellow | 331300 |
| Équipements | 12 000 | Basismodul: Wissenschaftliche Geräte | 7 000 |
| Frais de missions | 61 800 | Basismodul : Reisekosten | 34 600 |
| Autres dépenses | 37 500 | Basismodul : Sonstige Kosten | 15 000 |
| | | Basismodul : Publikationskosten | 2 250 |
| Prestations de service | 12 000 | Basismodul : Mittel für Gäste | 2 400 |
| Frais de gestion | 12 508 | | |
| | | Modul Projektspezifische Workshops | |
| | | Modul Öffentlichkeitsarbeit | |
| **Total France** | **325 208 €** | **Gesamtsumme Deutschland** | **392 550** |
| **Gesamtsumme** | **717 758€** | | |

**6. Autres moyens engagés pour la réalisation du projet**

**Non**

**7. Autres financements éventuels**

Aucun autre financement n'a été sollicité pour le présent projet auprès d'autres organismes de financement. Si tel était le cas, nous nous engageons à en informer sans délai l'ANR et la DFG.

## Annexe 1[2]
## Corpus multilingues

### 1.1. Corpus français

| Corpus | Période | Mots-occurrences | Type de corpus | Langue |
|---|---|---|---|---|
| Fiction (Emolex) | 1950-2010 | 15.970.000 | Textes littéraires | français |
| Presse (Emolex) | 2007-2008 | 108.740.000 | Textes journalistiques | français |
| Scientext français | 2000-2010 | 4.880.000 | Textes scientifiques | français |
| TOTAL: | | 129.590.000 | | |

### 1.2. Corpus anglais

| Corpus | Période | Mots-occurrences | Type de corpus | Langue |
|---|---|---|---|---|
| Fiction (Emolex) | 1900-2010 | 30.000.000 | Textes littéraires | anglais |
| Presse (Emolex) | 2007-2008 | 105.500.000 | Textes journalistiques | anglais |
| Scientifique | 1970-2005 | 30.000.000 | Textes scientifiques | anglais |
| TOTAL: | | 165.500.000 | | |

### 1.3. Corpus allemands

| Corpus | Période | Mots-occurrences | Type de corpus | Langue |
|---|---|---|---|---|
| Fiction (Emolex) | 1950-2010 | 6.370.000 | Textes littéraires | allemand |
| Presse (Emolex) | 2007-2008 | 93.700.000 | Textes journalistiques | allemand |
| TOTAL: | | 100.070.000 | | |

[2] Les corpus français, anglais et allemands dont nous disposons dans l'Emobase seront complétés et équilibrés.