Hétérogénéité et extraction d'information factuelle dans un corpus de récits de voyage

Anaïs Lefeuvre, Natalia Vinogradova

Hétérogénéité et extraction d'information factuelle dans un corpus de récits de voyage
Heterogeneity and factual information extraction in a corpus of travel writing

Anaïs Lefeuvre, Natalia Vinogradova, Université Bordeaux 1, INRIA SIGNES – LaBRI

Résumé
L'extraction d'information nécessite une connaissance des objets à extraire. Nous cherchons dans ce travail à décrire le comportement des séquences textuelles présentant l'itinéraire au sein du récit de voyage. Le récit de voyage est reconnu comme genre hétérogène, nous analysons donc cette hétérogénéité afin de pouvoir reconnaître les séquences homogènes, dont la description d'itinéraire fait partie. Nous menons notre analyse à plusieurs niveaux discursifs, ce qui nous permet d'avoir une vision globale du comportement de notre objet d'étude, l'itinéraire, et du contexte dans lequel il apparaît. Dans la perspective de l'extraction automatique d'itinéraire, nous utilisons de nombreux outils, chacun adapté au mieux au niveau d'analyse traité. En nous appuyant sur le cadre théorique de la SRDT (*Segmented Discourse Representation Theory*), dont nous montrons l'adéquation à l'étude, nous approchons le fonctionnement des descriptions des itinéraires, nous poussant à enrichir la méthode d'extraction afin de gérer l'hétérogénéité des unités discursives dans lesquels l'itinéraire est énoncé.

Mots-clés : hétérogénéité, récit de voyage, itinéraires, séquences textuelles, segments de discours, extraction d'information, SDRT et types fonctionnels.

Abstract
The information extraction task requires a good knowledge of the object to be extracted. In this work we explore the behavior of textual sequences describing the itinerary within the travel writing. Travel novel is a specific genre that is recognized to be heterogeneous, so we analyze its heterogeneity in order to discriminate homogeneous sequences, one of which being the itinerary description. Our analysis holds on different discourse levels, it allows us to get an overview of itinerary behavior through the narration. In order to automatize the extraction of itineraries, we use different tools, each one being perfectly adapted to the discourse level in question. Our theoretical framework at the semantic representation level, the SDRT (*Segmented Discourse Representation Theory*), complies with such kind of analysis, as we see in the course of this work. This study makes us understand the itinerary sequences behavior, leading us to enrich our extraction method to cope with heterogeneity of the discourse units dedicated to the itinerary.

Key words: heterogeneity, travel writing, itineraries, textual sequences, discourse segments, information extraction, SDRT and functional types.

1. Introduction

Les travaux présentés dans le cadre de cette contribution s'inscrivent dans le projet ITIPY, « Itinéraires Pyrénéens », qui a pour objectif l'extraction automatique des itinéraires dans un corpus de récits de voyage du XIXème siècle. Nous nous attachons à reconnaître, décrire et caractériser l'expression du déplacement dans notre corpus. Nos enjeux recoupent ceux de Enjalbert (2003) dans ses travaux sur la recherche et l'extraction de l'information spatiale. L'objet de ce travail est le récit de voyage (désormais RV) posé comme un genre discursif à part entière, mais marqué également par une hétérogénéité interne.

L'organisation du discours comme objet linguistique se fait à plusieurs niveaux. Nous distinguons tout d'abord l'organisation superficielle qui assure la cohésion entre les unités

discursives et l'organisation profonde qui garantit la cohérence du discours. L'analyse profonde du discours ne peut être obtenue que suite à l'analyse de chaque niveau discursif. Deux techniques peuvent être utilisées, « top-down » et « bottom-up ». Elles consistent soit à retrouver les spécificités discursives en s'appuyant sur la structure du discours, soit à observer les unités minimales de niveau inférieur pour ensuite reconstruire l'architecture du discours aux niveaux supérieurs. Nous verrons dans cet article que l'hétérogénéité des données discursives requiert une approche où les deux techniques sont complémentaires. Nous utilisons tout d'abord l'analyse « top-down», considérant en premier lieu la production discursive propre au récit de voyage, puis les différents récits et le type d'information qui est délivré dans chacun d'entre eux. Ensuite, nous observons la structure segmentale et séquentielle contenant l'information pertinente dans une logique « bottom-up ». Enfin, nous observons les segments au regard de l'itinéraire et indépendamment de la structure du discours. La Figure 1 présente comment les deux approches se rejoignent dans une analyse discursive visant à dégager les informations pertinentes.

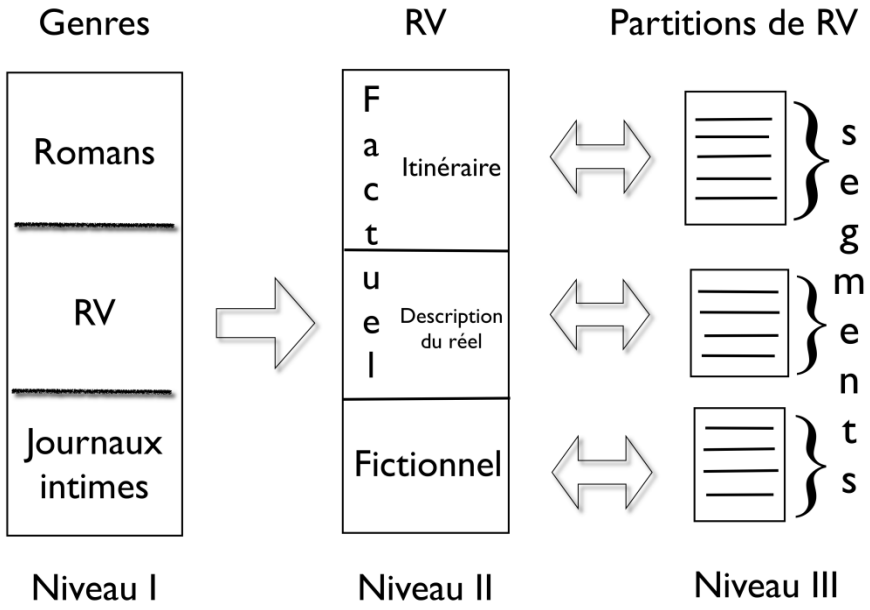

Figure 1. Schéma d'analyse par niveau

Suite au travail de Magri-Mourgues (2009), au premier niveau, nous cherchons à confronter notre corpus de RV avec des œuvres littéraires de genre similaire, puis avec des romans de la même période, et finalement avec des journaux intimes. Quelques marqueurs utilisés dans le travail de Magri-Mourgues pour dresser un portrait du RV seront testés afin d'en éprouver la fiabilité pour notre corpus. Nous supposons que l'hétérogénéité des données textuelles reconnue de ce genre doit se manifester lors de la confrontation à un autre corpus en variant les genres. Ensuite, nous approchons le corpus sous l'angle de la théorie de Pasquali (1995) selon laquelle le RV est une alternance de séquences textuelles comprenant « le récit de voyage et de découverte du réel », « le récit métaphorique », etc. (p. 94). Le RV présente une dichotomie explicite entre information factuelle, dans laquelle une place particulière est réservée aux itinéraires, et information fictionnelle, définie comme « types fonctionnels » par Biber et *al.* (2007). Dans le cadre de la SDRT (*Segmented Discourse Representation Theory*, Asher et Lascaridès 1993), nous poursuivons l'analyse au niveau inférieur et examinons les unités discursives minimales, ou segments, afin d'accorder une représentation sémantique au discours et de connaître le comportement discursif des itinéraires, que nous voulons extraire. À ce niveau, nous questionnerons une nouvelle fois l'hétérogénéité des données telle que nous

la définissons au cours de l'analyse, ce qui nous amènera à considérer l'extraction des segments obéissant à un patron lexico-syntaxique indépendamment de la structure séquentielle du RV.

2. Le récit de voyage et ses spécificités

2. 1. Le corpus et le genre discursif

Un corpus contrastif à partir d'un corpus de référence de RV a été constitué afin de varier les conditions de production discursive et de cerner au mieux le RV. Les principaux paramètres en jeu ici sont : le genre, l'opposition factuel-fictionnel, l'environnement spatial du récit, l'auteur, et les dates de production. Le corpus initial est composé de onze œuvres identifiées, par la médiathèque de Pau, comme récits de voyage pyrénéens du XIX^{ème} et début XX^{ème} siècles. Le RV est reconnu comme un genre caractérisé par des spécificités attestées dans *Le voyage à pas comptés* (Magri-Mourgues 2009). Nous avons utilisé quelques critères cités dans ce travail afin d'identifier le corpus de départ ITIPY comme appartenant au RV et de renforcer cette classification par notre exploration. Ainsi, le corpus IPITY a été augmenté de différents textes permettant de délimiter plusieurs corpus de travail (pour une présentation en détail du corpus constitué, un tableau récapitulatif se trouve en annexe 1):

- deux RV se déroulant en Normandie et en Orient, qui forment avec onze textes originaux de ITIPY la partition A ;
- sept romans dont trois du même auteur et quatre d'auteurs divers regroupant différents genres, allant du « roman personnel » (Dufiez-Sanchez 2010) au « roman naturaliste », ceux-ci composant la partition B ;
- trois journaux intimes datant de la fin du XIX^{ème} pour les deux premiers et de la fin du siècle précédent pour le dernier, constituant la partition C.

Nous espérions *a priori* faire ressortir les spécificités du RV par la confrontation à des œuvres ayant certains paramètres de production similaires. En effet, il ressort de la comparaison globale de ces partitions que le RV se distingue des autres genres du roman par son régime factuel. L'univers du discours n'est donc plus décroché de la réalité extralinguistique mais se proclame au contraire « en phase » avec cette réalité pour que l'authenticité présumée du récit n'en soit que plus frappante. Cette volonté d'embrayer l'univers du discours avec une réalité extralinguistique fait bien partie des conditions de production du RV et en influence la réalisation énonciative.

2. 2. La construction de l'univers discursif

En nous appuyant sur l'analyse de Magri-Mourgues (2009), nous avons procédé à une étude lexicométrique à l'aide du logiciel *Lexico3*[1] (Salem 1991) afin d'observer et de défricher notre corpus dans son ensemble. *Lexico3* est destiné à opérer de manière quantitative une analyse basée sur les lois d'hypergéométrie des discours en vue d'une approche qualitative ultérieure. Sans surprise, les champs lexicaux ayant trait aux éléments naturels comme « montagnes », « vallée », « rivière », par exemple, sont principalement attachés à la partition A, tandis que les objets quotidiens et les parties du corps, « tête », « yeux », « mains », sont plutôt attachés à la partition B. On notera tout de même le cas particulier de « pied(s) » qui se trouve en fréquence supérieure dans la partition A du fait des constructions syntagmatiques « au pied de la montagne » et « sous leurs pieds ». L'univers est spatialisé, et l'expérience narrée est celle de l'énonciateur face au territoire, à la nature, tandis que dans le roman, l'aventure narrée se situe plus particulièrement dans la société.

L'analyse factorielle des correspondances fournie par *Lexico3* montre un regroupement

[1] http://www.tal.univ-paris3.fr/lexico/

des textes de récit de voyage (en noir), nettement distingués de la partition B (blanc) et un rapprochement de la partition C (gris) vers le RV.

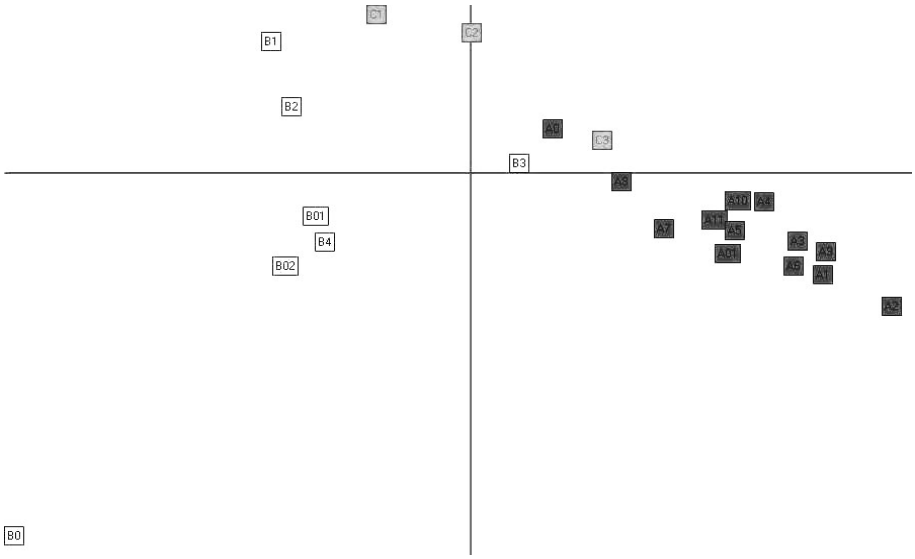

Figure 2. Analyse factorielle des correspondances

La spécificité des RV est ainsi confirmée. Elle est notamment illustrée par la répartition de certains marqueurs énonciatifs, tout particulièrement les indices de la deixis et l'expression de la personne. Chacune des partitions traitées, malgré une part d'hétérogénéité propre, peut dès lors recevoir une caractérisation linguistique.

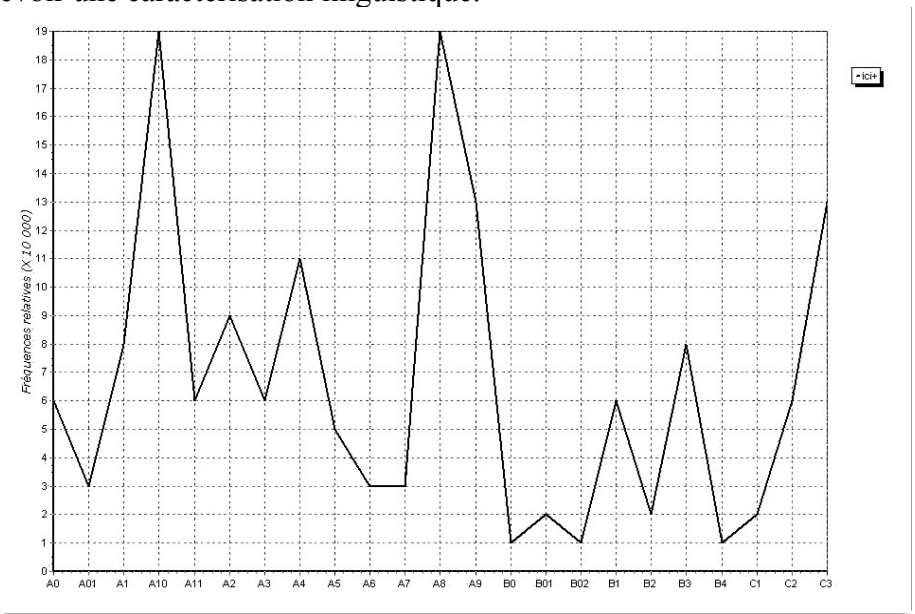

Figure 3. Fréquences d'« ici »

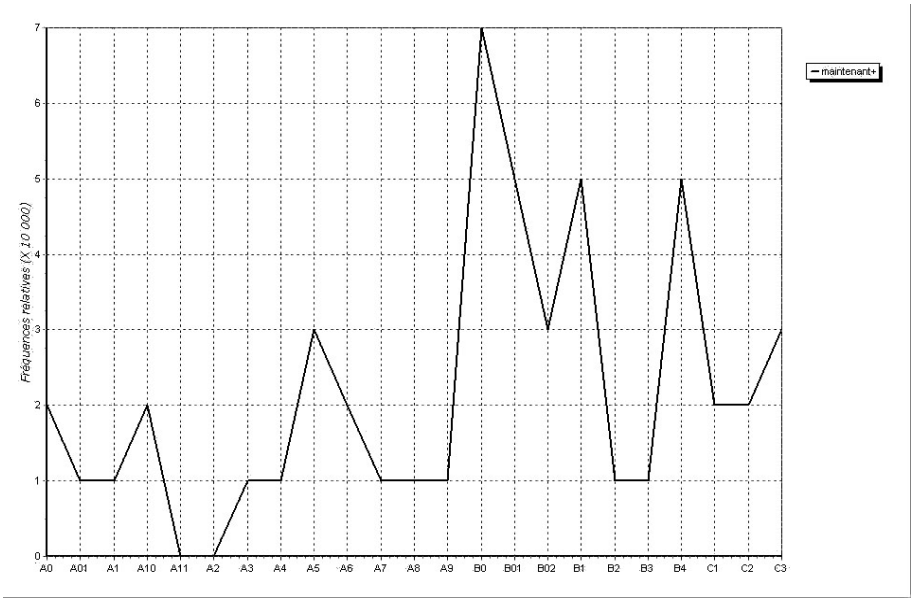

Figure 4. Fréquences de « maintenant »

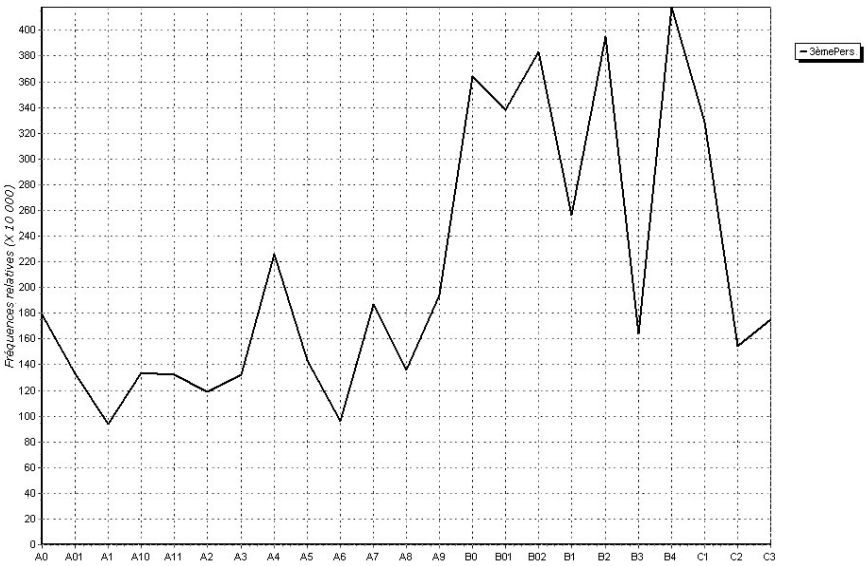

Figure 5. Fréquences du groupe de forme 3ème pers. Sg., pl., masc. et fém.

Les figures 3 et 4 représentent la ventilation fréquentielle des adverbes « ici » et « maintenant », afin d'illustrer le brayage spatial et temporel grâce à des formes morphologiquement simples et sans équivoque. Nous donnons ici ces fréquences à titre indicatif ainsi que la figure 5, qui présente une répartition discriminante de la 3ème personne. Le brayage spatial, primant nettement sur le brayage temporel pour la partition A, exprime une caractérisation du RV malgré des fréquences restant très hétérogènes. Le brayage évoqué par les personnes suit la même logique. Le RV dénote un monde spatialisé, et dans lequel un « je » avance dans un espace décrit, sans nécessiter systématiquement le rappel du fil temporel de la narration. Par contraste, au sein du corpus, le roman comporte généralement un narrateur hétérodiégétique qui relate les faits et actions autour du protagoniste de la fiction. La spatialisation est moins prégnante et le rapport entre la deixis temporelle et spatiale est plus équilibré.

On observe des comportements semblables entre tous les RV et B3, C2, C3 (visible dans la figure 2). D'après les analystes du roman du XIXème, *René* de Châteaubriant (B3), est

catégorisé comme un « roman personnel » dont la portée déictique est franchement contrastée par rapport à ses contemporains. Nous prendrons en compte dans notre analyse le statut particulier de ce genre. C2 et C3 quant à eux, sont nos témoins. Leur comparaison avec les RV permet de les rapprocher de la partition dite *factuelle* sans que pour autant, ils ne fassent partie des récits de voyage : ils partagent avec les RV un aspect de « découverte du réel », le réel étant un territoire, et la volonté de relater des faits authentiques.

Nous constatons que le RV est un genre spécifique qui possède des caractéristiques lexicales et déictiques attestées. Nous remarquons néanmoins les similarités avec les textes des autres genres et des fréquences hétérogènes d'un texte à l'autre. Ce fait nous invite à explorer notre corpus d'origine de plus près, en envisageant notamment le niveau d'analyse intra-textuel afin de mieux comprendre son organisation. Ainsi, nous nous attachons à l'étape suivante de notre analyse, à observer au sein du RV, c'est-à-dire de la partition A, les manifestations de l'alternance des différents récits. Malgré les spécificités observées caractérisant la partition A, nous avons déjà remarqué une forme d'hétérogénéité dans la construction de l'univers discursif de cette partition, nous nous intéressons maintenant à la construction interne au RV, représentée par le niveau II de la figure 1.

3. L'hétérogénéité et les séquences textuelles

3. 1. Les types fonctionnels

Selon Adam (1993, en ligne) tout texte est « une configuration réglée par divers modules ou sous-systèmes en constante interaction », l'un de ces modules étant son organisation séquentielle. La séquence est définie comme « une entité textuelle constituée de paquets de propositions (les macro-propositions), elles-mêmes constituées de n propositions ». Le RV, tout comme les autres productions discursives, est composé de séquences textuelles, que l'on peut discriminer les unes des autres par divers critères. Suite à la théorie d'Affergan dans *Exotisme et altérité* en 1987, Pasquali (1995 : 94) souligne que « le récit de voyage et de découverte du réel [alterne avec] le récit métaphorique, le récit métonymique et le récit synecdochique ». Il existe donc une structuration propre au RV dans laquelle on trouve différents procédés narratifs. Il semble que cette hétérogénéité des séquences textuelles est intrinsèque au RV.

Pour traiter ce phénomène, nous adoptons la terminologie de Biber (1998) qui parle de *types fonctionnels* pour désigner l'ensemble des séquences textuelles ayant le même but communicatif. Le mécanisme de production du discours, qu'il soit oral ou écrit, est régi par un but communicatif : *donner une information*, *demander un renseignement*, *formuler un ordre*, etc. qui influence le type fonctionnel. Dans le discours, ces buts communicatifs alternent, ce qui se traduit par une alternance des types fonctionnels. Nous remarquons ce même phénomène au sein du RV qui se prête tout à fait à une approche thématique factuel/fictionnel. Biber a construit cette catégorisation de manière interne au texte, en croisant les marqueurs linguistiques émergents. Ces marqueurs sont des traits lexicaux et grammaticaux qui permettent de caractériser les différents types fonctionnels et de les distinguer les uns des autres. Nous citerons en guise d'exemple la dimension « informative » *versus* « impliquée » regroupant plusieurs traits tels que les pronoms personnels employés, par exemple : l'utilisation de « je » proposant plutôt un discours impliqué et l'absence du locuteur présentant plutôt un discours informatif (Pery-Woodley 1994). Dans le RV, ces deux dimensions se présentent par le récit factuel et le récit fictionnel respectivement. En effet, dans le récit factuel, plutôt informatif, l'énonciateur prétend décrire le monde réel comme il est, tandis que dans le récit fictionnel, plutôt brayé (impliqué), il cherche à montrer des émotions, et une forme de subjectivité. Puis, au sein du récit factuel, un découpage plus fin nous intéresse, à savoir le type fonctionnel propre au récit de l'itinéraire qu'il faudra séparer du récit de la vie quotidienne ou encore des descriptions naturalistes.

3. 2. Spécificités des types fonctionnels dans le RV

Dans notre corpus, certaines séquences textuelles régies par un type fonctionnel sont explicitement délimitées par le découpage en paragraphe. Prenons un exemple, extrait de *Fragments d'un voyage sentimental et pittoresque dans les Pyrénées* :

(1) Mais quittons ce lieu désolé. M. Mercére mon protecteur, part pour Gavarnies. Je vais l'accompagner à travers les hautes montagnes qui séparent Héas de ce dernier district. Quelle route, grands Dieux ! Mais que dis-je ? Il n'y a point de route ici : le voyageur monte, descend, traverse les prairies & les gaves, sans chemins, sans traces, sans autre renseignement, que la position respective du lieu d'où il vient, & du lieu où il va. M. Pasumot, M. Dusaulx, & vous curieux, amateurs, ou promeneurs de Barège, je commence à vous féliciter de n'être point du voyage. Ne croyez point cependant que je forme le moindre regret de l'avoir entrepris ; les montagnards & les montagnes me seront plus connus désormais, que je n'aurais pu me flatter de les connaître ; dans cinq cents de nos courses ordinaires. Mais revenons, partons de Héas. [A10 : *Fragments d'un voyage sentimental et pittoresque dans les Pyrénées*].

Le contenu factuel des trois premières phrases, suivies par une expression forte de subjectivité, introduit un récit plutôt fictionnel. Nous notons une rupture fonctionnelle, marquée par des commentaires subjectifs sur la difficulté du voyage puis la clôture de cet aparté par « Mais revenons ». Ce passage indique clairement que nous nous situons sur deux plans énonciatifs liés : le premier décrivant le voyage réalisé dans le passé, ancré dans une réalité révolue, et le second incluant le lecteur dans le voyage qui se déroule au fil du récit. Ce passage relève du même phénomène que nous avons décrit avec l'exemple des pronoms avec *Lexico3*. Le « nous » inclusif permet de délivrer des informations factuelles sur l'itinéraire, tout en brayant l'univers discursif et en intégrant le lecteur. L'aparté central dissocie le protagoniste du lecteur par l'opposition « je/vous », et laisse place à des informations d'ordre fictionnelle.

Le repérage de ces types fonctionnels laisse apparaître les traces de l'hétérogénéité structurelle et énonciative du RV. Trois types fonctionnels sont pertinents pour notre recherche : le récit fictionnel teinté d'une forte subjectivité de l'énonciateur, le type factuel de l'itinéraire, incluant une deixis personnelle et spatiale, plus faible en subjectivèmes lexicaux, et le récit factuel « autre », englobant les descriptions de l'environnement et le récit de la vie quotidienne.

Nous rappelons que nous nous situons à la seconde étape de l'analyse (allant du niveau II au niveau III de la figure 1) au sein de laquelle nous voulons dégager les séquences textuelles pertinentes à l'extraction des itinéraires. Ainsi, nous pouvons conclure que l'hétérogénéité des données doit être prise en compte dans l'extraction d'information. Introduisons maintenant le cadre théorique dans lequel nous représentons la sémantique du discours (désormais nous nous acheminons du niveau III au niveau II).

4. Du segment au discours

4. 1. La segmentation du discours dans le cadre de la SDRT

Dans le cadre théorique de la SDRT (*Segmented Discourse Representation Theory*), version enrichie de la DRT, nous cherchons à traiter automatiquement les informations factuelles ayant trait aux itinéraires. Dans ce formalisme logique, on dresse un univers du discours dans lequel l'énoncé est vrai. Le discours est défini comme une unité sémantique contextuelle, de laquelle on dégage d'une part les entités présentes dans l'univers du discours, et d'autre part les conditions de vérité de cet univers. A l'aide d'un lexique et de l'analyse syntaxique, on

extrait du discours toutes les informations permettant de mettre en relief les propriétés de chaque entité, et les relations entre ces entités qui sont vraies dans ce modèle. On représente généralement l'univers par une boîte qui prend la forme suivante (Amsili 1998) :

(2) Pedro possède un âne.

x y
Pedro(x) âne(y) posséder (x,y)

Figure 5. DRS de « Pedro possède un âne »

La SDRT permet d'intégrer les relations rhétoriques entre segments au sein du discours. L'unité utilisée est l'EDU *elementary discourse unit*, définie comme proposition élémentaire, comme présentée dans l'exemple (2). Autrement dit, l'EDU correspond à un segment textuel minimal du discours représentant soit un événement, soit une période temporelle. Pour se rapprocher de l'analyse textuelle, on dira qu'une phrase peut être composée de plusieurs segments. On peut aussi vouloir relier un bloc de plusieurs segments à un seul segment par une relation, on parlera alors de CDU, *complex discourse unit* pour désigner le bloc. Les CDU contiennent plusieurs unités élémentaires, en préservant la cohérence thématique et rhétorique. La définition formelle de ce concept est proposée dans le travail de Asher, Venant, Muller et Afantenos (2011). Nous proposons un exemple inspiré de notre corpus et analysé en SDRT :

(3) [a] Notre arrivée à Bagnères ne fût pas facile.
- [b] Nous avancions à un rythme irrégulier depuis un moment
- [c] et nous entrâmes tard dans la rue Saint-Blaise.
- [d] Depuis là-bas, nous fîmes une promenade à Superbagnères.
- [e] Nous suivions d'abord un chemin d'une pente peu rapide
- [f] et ensuite nous nous élevâmes presque sans nous en apercevoir jusqu'à Superbagnères.

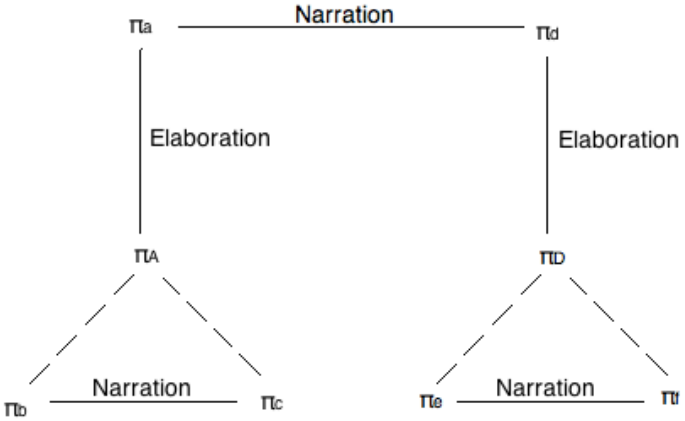

Figure 6. Analyse en SDRS

Les EDU [a], [b], [c], [d], [e] et [f] sont des boîtes comme présentées précédemment, que l'on traduit par des labels. Respectivement π_a, π_b, π_c, π_d, π_e et π_f .π_A et π_D sont les labels de CDU, contenant π_b et π_c pour le premier, et π_e et π_f pour le second. Ici π_a est élaboré par π_A et π_d est élaboré par π_D.

L'intégration des relations discursives à la DRT a été motivée par l'impossibilité d'exprimer la temporalité d'une suite d'événements en s'appuyant simplement sur la syntaxe et sur le lexique, et par la nécessité de faire appel à des connaissances du langage et du monde pour une représentation sémantique du discours juste. En effet, la relation *elaboration* entre π_a et π_A n'est pas donnée par la syntaxe, et seule la connaissance de la géographie de Bagnères et de ses rues permet de s'assurer que π_A contenant l'entrée dans la rue Saint-Blaise est une description des étapes utiles à l'arrivée à Bagnères. *Elaboration* étant définie[2] comme une relation subordonnante, permettant de relier les deux événements respectivement dénotés dans les segments π_a et π_b par une inclusion thématique et temporelle. La relation temporelle entre les trois évènements portés par π_a, π_b et π_c n'est pas une succession stricte portée éventuellement par une relation de *narration*, mais bien un détail des étapes, π_b et π_c, de l'événement porté par π_a, c'est pourquoi la relation discursive est *elaboration*, impliquant une inclusion temporelle au moins partielle de π_A dans π_a.. Les relations discursives *elaboration* et *narration* sont deux relations incompatibles, deux événements ne pouvant par ailleurs pas être successifs et inclus l'un dans l'autre. Nous présentons quelques unes des relations disponibles dans la section suivante.

4. 2. Segmentation et relations discursives dans le RV

Regardons de plus près notre corpus et reprenons l'exemple (1) segmenté en (4) :

(4) [Mais quittons ce lieu désolé.]π_a [M. Mercére mon protecteur, part pour Gavarnies.]π_b [Je vais l'accompagner à travers les hautes montagnes]π_c [qui séparent Héas de ce dernier district.]π_d
[Quelle route, grands Dieux !]π_e [Mais que dis-je ?]π_f [Il n'y a point de route ici :]π_g [le voyageur monte, descend, traverse les prairies & les gaves, sans chemins, sans traces, sans autre renseignement, que la position respective du lieu]π_h [d'où il vient]π_i [, & du lieu] $\pi_{h\ suite}$ [où il va.]π_i [M. Pasumot, M. Dusaulx, & vous curieux, amateurs, ou promeneurs de Barège, je commence à vous féliciter de n'être point du voyage.]π_k [Ne croyez point cependant]π_l [que je forme le moindre regret de l'avoir entrepris ;]π_m [les montagnards & les montagnes me seront plus connus désormais, que je n'aurais pu me flatter de les connaître ; dans cinq cents de nos courses ordinaires.]π_n [Mais revenons,]π_o [partons de Héas.]π_p

Nous analysons cet exemple en EDU et CDU et attribuons les relations entre les segments. Néanmoins, comme montré dans la section précédente, cet extrait est hétérogène, et contient des segments de deux types fonctionnels différents. A l'origine, la SDRT propose des relations entre les segments de même type fonctionnel. La question se pose alors de définir comment élargir le cadre théorique pour pouvoir analyser correctement cet extrait en particulier et le RV de manière plus générale. Voici l'analyse des relations discursives entre les EDU et CDU de l'exemple (4).

[2] Nous donnons ici des éléments de compréhension informels.

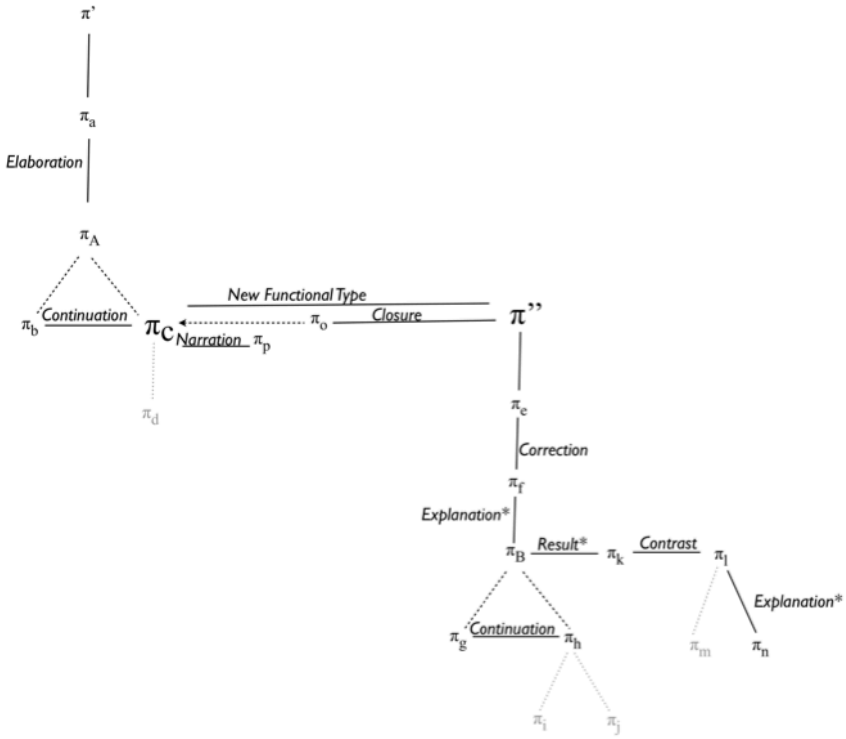

Figure 7. Analyse en SDRS

La séquence textuelle ayant trait à l'itinéraire semble caractérisée, dans ce passage, par les relations de discours telles que *elaboration* et *continuation* (π_a, π_b, π_c). Nous avons déjà introduit la relation *elaboration*. Quant à la relation *continuation*, relation coordonnante, elle implique une forme de narration sans domination thématique de l'une sur l'autre et sans succession temporelle entre les événements portés par les deux EDU. Les constituants de la seconde séquence sont liés par les relations de *correction*, permettant de remettre en question le contenu sémantique de son premier argument par le second, *continuation* (expliquée précédemment), *explanation** et *result** qui introduisent respectivement l'explication et le résultat meta-discursifs, et enfin *contrast*, permettant d'exprimer par le second segment la violation d'une attente induite par le premier. Nous précisons que nous ne nous intéressons pas pour le moment aux segments imbriqués, en gris sur le graphe.

Il faut maintenant établir la relation entre les deux séquences qui se suivent dans notre exemple. Afin de conserver la cohésion énonciative et la cohérence temporelle, il faut considérer que la seconde séquence est une forme de cadre (Charolles 2005), le cadre étant un regroupement de segments homogènes sous la portée d'un critère sémantique qui permet de structurer le discours. Notre critère pragmatico-sémantique portant sur l'itinéraire conserve une cohérence énonciative et temporelle.

La SDRT permet d'établir les relations entre les segments discursifs simples et complexes, elle est parfaitement adaptée à l'analyse des relations entre cadres discursifs. Pourtant, ce comportement nous semble ne pas avoir été traité au sein de la SDRT : quelle relation discursive utiliser entre nos deux séquences ? Aucune ne convient, il faut donc introduire une nouvelle relation permettant de manipuler le phénomène rencontré.

4. 3. Proposition d'une nouvelle relation discursive

La structure du RV repose principalement sur l'avancée de l'itinéraire et les autres types d'informations dépendent de cette structure. Ainsi, nous avançons que les informations intégrées dans le RV, sont rattachées au fil narratif de l'itinéraire sous la forme de cadres

discursifs. Chacun contient des informations fictionnelles ou factuelles spécifiques qui n'ont pas forcement de lien avec le cadre précédent. À la fin d'une séquence fictionnelle par exemple, le cadre se ferme et l'auteur poursuit avec la description de l'itinéraire. Ce retour à l'itinéraire, après un ou plusieurs cadres, est donc systématique, il garantit la cohérence au sein du RV. Dans l'exemple étudié, l'ouverture et la fermeture du cadre sont très explicites du point de vue énonciatif et des relations discursives.

Nous proposons deux relations rhétoriques généralisées propres à ce type de discours : *New Functional Type* et son dual *Closure*. En effet, les informations fictionnelles étant variées, nous supposons plusieurs types fonctionnels correspondants. Dans ces conditions, chaque nouveau cadre discursif est régi par le type fonctionnel des informations qu'il contient. Lorsque le cadre se ferme, la relation de clôture renvoie au moment où la description de l'itinéraire s'est arrêtée.

Dans la SDRT, il est reconnu que la liste des relations n'est pas complète et que les relations sont souvent ambiguës. Nous cherchons à enrichir le cadre théorique en ajoutant deux nouvelles relations à cette liste, ce qui nous permettra de traiter le phénomène décrit. Notre proposition est le fruit d'une nécessité, elle se restreint à un phénomène spécifique à un seul genre discursif et beaucoup de problèmes restent encore à résoudre[3].

Néanmoins, on ne peut généraliser totalement le fonctionnement textuel de l'itinéraire, car nous pouvons prévoir de nombreux cas où ce schéma ne s'applique pas. Prenons l'exemple suivant :

(6) En faisant ces réflexions désolantes pour moi qui désirerais bien vivement partager avec mes amis tout le plaisir que je prends, toutes les sensations que j'éprouve, **nous arrivâmes dans le district de Grip** où M. l'abbé P*** ne place qu'une seule maison, où nous ne vîmes qu'un seul village : cela saute aux yeux. Comment se peut-il qu'on n'indique en cet endroit qu'une seule, qu'une unique maison, lorsque la route & la vallée entières sont couvertes d'habitations, depuis un quart de lieue au-dessus de l'auberge jusqu'au village de Sainte-Marie ? [A10 : *Fragments d'un voyage sentimental et pittoresque dans les Pyrénées*]

On voit que le segment appartenant au type fonctionnel de l'itinéraire est isolé dans une séquence dédiée à l'expression des sentiments du protagoniste, puis à la description de son environnement. Le changement de type fonctionnel n'est pas explicite et, comme précédemment, aucun marqueur ne le réalise. Pour autant, il nous faut relier la représentation de ce segment à la suite de la représentation dédiée à l'itinéraire. La conclusion que nous pouvons tirer de l'analyse de cet exemple est que nos séquences textuelles sont effectivement hétérogènes.

Notre analyse du RV montre que l'itinéraire ne peut être traité exclusivement en terme de cadres. Les données textuelles du RV étant hétérogènes, les segments sont parfois isolés et ceci nous informe que la structuration du discours en séquences ne peut être limitée à des cadres s'ouvrant et se fermant. En complément de ces observations et de la proposition de relation discursive, nous avons voulu analyser les diverses formes que pouvait revêtir le segment dédié à l'itinéraire. En restant au niveau III de la figure 1, nous avons testé une technique traditionnelle en extraction d'information pour être plus précises sur notre analyse de l'hétérogénéité dans le RV.

[3] Tels que le respect de la frontière droite que nous n'avons pas la place de développer ici (Asher 2003).

4. 4. L'hétérogénéité et le patron lexico-syntaxique du segment

Nous venons de décrire une structure du RV qui semblait régulière. Cependant nous avons présenté un exemple qui remet en question cette régularité. Visant l'extraction automatique des itinéraires, nous ne pouvons pas nous contenter de l'extraction des séquences textuelles du cadre correspondant. Nous avons besoin de trouver une spécificité intrinsèque à notre type fonctionnel qui nous permettra d'extraire tous les segments de ce type, qu'ils soient isolés ou non.

Biber (2007) indique que chaque type fonctionnel induit certains marqueurs lexico-syntaxiques. L'utilisation des verbes de déplacement étant une spécificité lexicale propre aux itinéraires, nous nous sommes intéressées aux constructions syntaxiques contenant ces verbes. Plus précisément, nous avons basé l'analyse sur un travail réalisé dans le cadre du projet ITIPY, par Loustau (2008) puis repris par Bessagnet et *al.* (2010). Les auteurs étudient le comportement du patron lexico-syntaxique [verbe de déplacement, préposition ?, syntagme candidat, ES], dans lequel la préposition peut être présente ou non[4], et pour lequel « ES » indique une entité spatiale[5]. Ce patron est une schématisation minimale du segment et nous pensons pouvoir l'utiliser pour détecter les segments pertinents pour la description d'itinéraire.

Ce patron présente l'avantage d'être robuste. Nous reprendrons en exemple les segments « partons de Héas » (1) et « nous arrivâmes dans le district de Grip » (6) qui obéissent au patron décrit, que le cadre discursif précédant soit fermé ou non. Cependant, ce patron ne couvre pas la totalité des expressions de déplacement présentes dans le corpus. Selon l'évaluation de Loustau (2008), le patron ne représente que de 63% à 85% de ces expressions au sein du corpus et l'algorithme d'extraction automatique utilisant le patron permet d'extraire autour de 80% d'entre elles. Pour éprouver la validité du patron, nous testons sa distribution au sein du corpus afin de le confronter dans les différentes productions discursives représentées. Nous avons choisi d'extraire les concordances de 4 verbes. Plus précisément, nous avons choisi deux verbes dont l'emploi est polysémique, « arriver » et « passer », et deux verbes dont la sémantique propose clairement un déplacement dans l'espace, « parcourir » et « descendre ». Par ailleurs, ces deux derniers verbes se distinguent l'un de l'autre par l'emploi métaphorique du second, comme par exemple dans *Madame Bovary* :

(7) [...] elle l'avait descendu tout au fond de son cœur.

Cette utilisation métaphorique est prise en compte par la présence de l'entité spatiale dans les arguments du verbe de déplacement.

Nous remarquons la présence intéressante dans la partition A de la pronominalisation de l'objet quand il est spatial.

(8) Cette route que je viens de parcourir [...] a souvent sur son côté gauche de fort jolies petites maisons [...]. [A0 : *Mémoires d'un touriste*]

Ici, la position du constituant portant le déplacement dans la relative marque une cohésion forte entre les différents segments du discours. La distinction entre segment du type fonctionnel itinéraire et du type fonctionnel description est donc brouillée syntaxiquement au sein du même segment. En SDRT, on trouvera donc une légitimité à représenter les segments

[4] On notera que « je quitte Pau » ne contient pas de préposition.

[5] Seul les noms propres sont assimilés aux ES pour le moment. Précisons par ailleurs la nécessité de disposer d'un lexique de toute entité pouvant être spatialisée. Par exemple, « quitter l'université », extrait du corpus, peut aussi bien représenter l'institution dans l'explication d'un choix de carrière que le bâtiment dans la narration du quotidien d'un personnage.

imbriqués. L'hétérogénéité des réalisations lexico-syntaxiques est à prendre en compte en vue d'une application efficace qui capturerait toutes les étapes de l'itinéraire d'un RV.

L'emploi du verbe n'est donc pas restreint au seul sens locatif et l'expression de l'itinéraire par le verbe ne se réalise pas selon un seul patron lexico-syntaxique. L'utilisation alternative des deux approches pour cerner notre objet, du segment au discours, et du discours au segment, est légitime. Certains cas ne sont pas traitables par notre analyse dans le cadre de la SDRT, tels que les segments isolés, et d'autres ne sont pas traitables par l'extraction en patron, comme par exemple, l'anaphore spatiale adverbiale, la résolution de l'anaphore spatiale nominale, le rejet du patron dans la relative. Nous rappelons la nécessité de prendre en considération plusieurs niveaux d'analyse tels que le défend Swales (1990) afin d'aborder l'hétérogénéité des données discursives et textuelles. Les différentes techniques présentées sont en adéquation si l'on considère les formes d'hétérogénéité rencontrées dans notre corpus.

Conclusion

Nous avons cherché à dresser le portrait énonciatif du RV pour comprendre en corpus le comportement des séquences présentant l'itinéraire. Le corpus d'origine a été confronté à d'autres textes ayant des traits similaires. Nous avons démontré qu'en dépit de cette similarité la partition des RV se distingue des autres partitions en révélant ainsi une forme d'homogénéité et d'hétérogénéité à différents niveaux. L'analyse de l'hétérogénéité interne du RV a été abordée sous l'angle de sa structuration : elle laisse une place à l'existence de nombreux cadres discursifs de types fonctionnels différents, mais les organise de manière bien spécifique, en les greffant à la narration de l'itinéraire.

La SDRT dispose d'outils appropriés pour traiter le discours, néanmoins nous proposons une relation discursive supplémentaire afin de traiter ce phénomène. Les relations ouvrant et fermant un cadre fonctionnel nécessitent des définitions formelles que nous n'avons pas l'opportunité de développer ici. Finalement, cette analyse rend compte des cadres de narration de l'itinéraire au sein du RV. Sachant que les productions discursives sont de fait hétérogènes dans leur structure, nous avons cherché et trouvé des exemples où la structuration du RV, comme nous l'avons décrite, n'est pas respectée : le cas où un segment d'un type fonctionnel est isolé au sein d'une séquence d'un autre type.

Nous avons donc éprouvé un patron lexico-syntaxique rigide qui permet de détecter les segments de l'itinéraire dans n'importe quel cadre discursif. Néanmoins, certains problèmes apparaissent dans l'analyse. Tout d'abord, le patron ne couvre pas tous les segments de l'itinéraire. Nous pensons par ailleurs à la possibilité d'enrichir cette approche hybride grâce à d'autres tests, ce qui nous reste à explorer dans de futurs travaux.

Nous dirons en dernier lieu que l'hétérogénéité des données discursives fait parti des spécificités du RV au niveau de la structuration fonctionnelle du discours, de sa construction énonciative et de la construction des segments appartenant au récit de l'itinéraire, malgré la convergence de plusieurs conditions de production laissant penser que ces segments formeraient un ensemble plutôt homogène. On peut dire que l'hétérogénéité du RV à tous les niveaux, du discours dans son ensemble à la composition lexico-syntaxique des segments le structurant, légitime une mobilisation de plusieurs techniques afin d'extraire efficacement et la totalité des formes exprimant l'itinéraire.

Bibliographie:

ADAM J.-M. (1993), « Le texte et ses composantes », Semen 8. [http://semen.revues.org/4341]

AMSILI P. & BRAS M. (1998), « DRT et compositionnalité », Traitement Automatique des Langues 39 (1), 131-160.

ASHER N. & LASCARIDES A. (2003), Logics of Conversation, Cambridge: Cambridge University Press.

ASHER N.et al. (2011), "Complex Discourse Units and their Semantics", Proceedings of Constraints in Discourse (CID 2011), Agay-Roches Rouges: France. [http://passage.inria.fr/cid2011/lib/exe/fetch.php/cid2011_submission_9.pdf]

BESSAGNET M.-N. et al. (2010), « Extraction automatique d'un lexique à connotation géographique à des fins ontologiques dans un corpus de récits de voyage », TALN 2010 – 17e Conférence sur le Traitement Automatique des Langues Naturelles (19-23 juillet 2010), Montréal : Québec. [http://hal.inria.fr/hal-00536083/en]

BIBER D., CONNOR U. & THOMAS A. (2007), Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure, Amsterdam: John Benjamins Publishing.

BIBER D., CONRAD S. & REPPEN R. (eds) (1998), Corpus linguistics: Investigating language structure and use, Cambridge: Cambridge University Press.

CHAROLLES M. & VIGIER D. (2005), « Les adverbiaux en position préverbale : portée cadrative et organisation des discours », Langue française 148, 9-30.

DUFIEZ-SANCHEZ V. (2010), Philosophie du roman personnel de Chateaubriand à Fromentin 1802-1863, Genève : Droz.

ENJALBERT P. (2003), « Qu'est-ce qu'un corpus homogène ? Réflexions à partir d'expériences en extraction et recherche d'information », in G. Williams (éd.) (2007), Texte et corpus : Actes des Troisièmes Journées de la Linguistique de Corpus, Lorient : Université de Bretagne Sud, 227-238.

LOUSTAU P. (2008), Interprétation automatique d'itinéraires dans des récits de voyages D'une information géographique du syntagme à une information géographique du discours, Thèse de doctorat de l'Université de Pau.

MAGRI-MOURGUES V. (2009), Le voyage à pas comptés. Pour une poétique du récit de voyage au XIXe siècle, Paris : Honoré Champion.

PASQUALI A. (1995), « Récits de voyage et critique : Un état des lieux », Textyles 12, 21-32.

PÉRY-WOODLEY M.-P. (1994), « Une pragmatique à fleur de texte : Marques superficielles des opérations de mise en texte », in S. Moirand et al. (éds), Parcours linguistiques de discours spécialisés, Bern : Peter Lang, 337-348.

SALEM A. (1991), Lexico1 – Logiciel pour l'analyse lexicométrique des textes, E.N.S. de Fontenay- Saint-Cloud, 10 p.

SWALES J. (1990), Genre analysis: English in academic and research settings, New York: Cambridge University Press.

Annexes:

Annexe 1. Constitution du corpus

ID	titre	auteur	Fact v/s Fict	date	genre
A0	Mémoires d'un touriste	Stendhal	Factuel	XIX	Récit de voyage
A01	Voyage en Orient	Alphonse de Lamartine	Factuel	XIX	Récit de voyage
A1	Ascension au Pic du Nethou	Platon de Tchihatcheff	Factuel	XIX	Récit de voyage
A2	Au Pays des Isards	Les frères Cadier	Factuel	Début XX	Récit de voyage
A3	Excursions autour du Vignemale	Alphonse Meillon	Factuel	Début XX	Récit de voyage
A4	La conquête du Mont-Perdu.	Louis Ramond de Carbonnières	Factuel	XIX	Récit de voyage
A5	L'aventure du Vignemale	Didier Lacaze	Factuel	XIX	Récit de voyage
A6	Les Pyrénées ou voyages pédestres (Livre I : Béarn-PaysBasque)	Vincent de Chausenque	Factuel	XIX	Récit de voyage
A7	Ann Lister, première ascension du Vignemale	traduction de Luc Maury	Factuel	XIX	Récit de voyage
A8	Voyage aux Pyrénées	Hippolyte Taine	Factuel	XIX	Récit de voyage
A9	Voyage au Mont-Perdu et Dans la partie adjacente des Hautes-Pyrénées	Louis Ramond de Carbonnières	Factuel	XIX	Récit de voyage
A10	Fragments d'un voyage sentimental et pittoresque dans les Pyrénées.	Jean Florimond Boudon de Saint-Amans	Factuel	XIX	Récit de voyage
A11	Voyages inédits dans les Pyrénées	Alain Bourneton	Factuel	XIX	Récit de voyage
B0	L'assommoir	Emil Zola	Fictionnel	XIX	Roman naturaliste
B01	L'argent	Emil Zola	Fictionnel	XIX	Roman naturaliste
B02	La curée	Emil Zola	Fictionnel	XIX	Roman naturaliste
B1	Les trois mousquetaires	Alexandre Dumas	Fictionnel	XIX	Roman historique
B2	La petite Fadette	George Sand	Fictionnel	XIX	Roman social
B3	René	François –René Chateaubriant	Fictionnel	XIX	Roman du moi
B4	Madame Bovary	Flaubert	Fictionnel	XIX	Roman réaliste
C1	Journal	Jules Renard	Témoin	Fin XIX	Roman réaliste / Journal intime
C2	Mon Journal	Léon Bloy	Témoin	Fin XVIII	Roman réaliste / Journal intime
C3	Voyage en France	Arthur Young	Témoin	Fin XVIII	Roman réaliste / Journal intime