Discursive analysis of itineraries in a historical regional corpus of travels: syntax, semantics, and pragmatics in a unified type theoretical framework

Richard Moot*
Université de Bordeaux
LaBRI(CNRS)
SIGNES(INRIA)
moot@labri.fr

Laurent Prévot
Université de Provence
Laboratoire Parole et Langage
laurent.prevot@lpl-aix.fr

Christian Retoré*
Université de Bordeaux
LaBRI(CNRS)
SIGNES(INRIA)
retore@labri.fr

1 Introduction

In this paper we will discuss the application of (Segmented) Discourse Representation Theory and the Generative Lexicon to the analysis of a historical French corpus of itineraries in the Pyrénées. Our research will focus in particular on how type coercion (Pustejovsky, 1995) can help us give a correct analysis of cases of so-called "fictive motion" (Talmy, 1999), which is evident is phrases like.

(1) The road runs along the coast for two hours.

(2) The path descended abruptly.

This case is particular in that an entity (which is considered immobile and which, in the context, defines a *path*) is the subject of a movement verb and that the combination is interpreted as a generic statement about the nature of this path, without any movement necessarily taking place.

2 Context

The context of the current research project is to provide a semantic representation the paths traversed or described by the authors of the different books in a corpus of itineraries in the Pyrénées mountains. It is a 19th century corpus consisting of 576.334 words, containing (among many other details and descriptions) narratives of the routes followed and the places visited by the authors. Other information about the corpus and its spatio-temporal analysis can be found in (Loustau, 2008; Asher et al., 2008).

The problem on which we will focus here is that in many cases, non-agentive movement verbs, such as "descendre" (corresponding to *go/come down* in English) can take as its subject not only a person (sentence (3) below), but also an immobile subject, provided it can be associated (or *coerced*) to a path, such as a fence or a road (sentence (4)).

(3) Jean descend.

(4) Le chemin descend.

The phenomenon evident in Examples (1), (2) and (4) has been called "fictive motion", (Talmy, 1999): our mind's eye or an imaginary traveler moves along a stationary object such as a road, a fence or a table. It is our goal to give an account of this phenomenon in a type-theoretic framework.

The examples below, which are taken from our corpus, show some further particularities of fictive motion.

(5) Nous coupons ici un sentier qui vient du port de Barroude (...)
Here, we cross a path which comes from the pass of Barroude

(6) La route suit le gave qui vient de Gavarnie.
The road follows the mountain stream coming from Gavarnie.

(7) Plus loin, de nobles hêtres montent sur le versant (...)
Further away, noble beeches climb the slope

(8) (...) cette route qui monte sans cesse pendant deux lieues
this road which climbs incessantly for two miles

*Richard Moot and Christian Retoré have benefited from a grant from the Conseil Régional d'Aquitaine, in the context of the Itipy project and like to thank the other participants of the Itipy project for their discussion on many of the themes of this paper

(9) Le chemin pavé de calcaire et de pierres luisantes (...) serpente à travers fourrés de buis et de noisetiers
The road paved with limestone and shining stones winds across buxus and hazels shrubbery

In example (5), it is clear — by the meaning of "couper" (cross) — that the authors do *not* take the path described. In example (6) the path is indicated as following a river, while being unclear (outside of the larger context of the narrative) about whether or not the authors take this path. In example (7), there is no real path specified and we interpret the sentence as the author's gaze following a path along the beeches up the slope, whereas in example (8) there is adverbial modification: "incessantly" and "for two miles". Though the class of permitted adverbs is semantically restricted[1], it is possible to have temporal adverbs such as "the road went along the coast for two hours" and some manner adverbs such as "the path descended abruptly/slowly" which do not commit us to inferring that the author actually took the road. All of this suggests that we can interpret a static object as the (abstract) *process* of traveling along it.

Finally, example (9) shows that we can refer to both the physical aspect of the path (the stones from which it is built) and the itinerary which corresponds to traversing this physical path at the same time.

3 Syntax and Semantics

Our semantic framework is integrated in a wide-coverage categorial grammar for French (Moot, 2010a), which has been semi-automatically extracted from the French Treebank (Abeillé et al., 2003). The wide-coverage grammar and the Grail parser (Moot, 2010b) combine to parse unseen sentences with a precision comparable to the best-known categorial parsers for English.

Categorial grammars are an especially appropriate choice in the current context because of their tight integration of syntax and semantics: each derivation in a categorial grammar corresponds to a typed lambda term and this makes writing Montague-style semantics for categorial grammars particularly simple (see (van Benthem, 1987; Moot and Retoré, 2011)).

As is well-know, however, the possibilities of lambda-calculus semantics permit an integration with more modern theories of semantics, such as discourse representation theory (Kamp and Reyle, 1993; Muskens, 1994) and the Generative Lexicon (Pustejovsky, 1995; Asher, 2011).

(Bos et al., 2004) and (Moot, 2010b) show that wide-coverage compositional semantics using DRT is possible (for English and for French respectively). In addition (Bassac et al., 2010; Mery, 2011) show that ideas from the Generative Lexicon can be implemented using polymorphic lambda term assignments to lexical entries.

The current paper proposes an extension to the work of (Moot, 2010b), along the lines proposed by (Bassac et al., 2010; Mery, 2011), which permits the system to handle cases of fictive movement.

4 The Type-Theoretic Framework

As (Bassac et al., 2010; Mery, 2011), we use $\Lambda\mathsf{Ty}_n$ as our type-theory. $\Lambda\mathsf{Ty}_n$ is the second-order lambda calculus with n sorts. It is multi-sorted in the sense that it has (in addition to the type t for truth values) more than a single type of entities e (see (Muskens, 1996; Moot and Retoré, 2011)). It is second-order in the sense of Girard's system F (see (Girard et al., 1988)): it allows us to quantify over types.

Definition 1 (Types) *The* types *of* $\Lambda\mathsf{Ty}_n$ *are.*

- *basic types,* t *for truth values,* v *for events, as well as several types corresponding to the different kinds of individuals in our domain, which includes* paths*,* physical objects*,* humans*, etc.*
- *type variables* α*,* β*,* $\ldots$
- *if* T_1 *and* T_2 *are types, then* $T_1 \rightarrow T_2$ *is a type.*
- *if* T *is a type and* α *is a type variable (not necessarily occurring in* T*), then* $\Pi\alpha.T$ *is a*

[1] In English we can use manner of motion verbs such as "crawl" and "run" as well

type, we will call types containing subtypes of the form $\Pi\alpha.T$ polymorphic *types.*

Definition 2 (Terms) *The* terms *of* $\Lambda\mathsf{Ty}_n$ *are.*

- *a variable* x^T *of type* T *is a term of type* T*.*
- *a constant* c^T *of type* T *is a term of type* T*.*
- $(f\ \tau)$ *is a term of type* T *if* f *is a term of type* $U \to T$ *and* τ *is a term of type* U *(application).*
- $\lambda x^U \tau$ *is a term of type* $U \to T$ *if* x *is a variable of type* U *and* τ *is a term of term of type* T *(abstraction).*
- $\tau\{U\}$ *is a term of type* $T[\alpha := U]$ *if* τ *is a term of type* $\Pi\alpha.T$ *and* U *is a type (universal application).*
- $\Lambda\alpha.\tau$ *is a term of type* $\Pi\alpha.T$ *if* α *is a type variable which does not occur freely in the type of a free variable and if* τ *is a term of type* T *(universal abstraction).*

Definition 3 (Conversion) *The* conversions *of* $\Lambda\mathsf{Ty}_n$ *are the following.*

- $(\lambda x.\tau)u$ *reduces to* $\tau[x := u]$ *(with the usual condition that no free variable in* u *is bound in* $\tau[x := u]$*, which we can always guarantee by renaming variables)*
- $(\Lambda\alpha.\tau)\{U\}$ *reduces to* $\tau[\alpha := U]$ *(remember that* α *and* U *are types).*

Definition 4 (Lexical Entries) *A* lexical entry *for a word* w *in* $\Lambda\mathsf{Ty}_n$ *is defined as a term of* $\Lambda\mathsf{Ty}_n$ *which is the base term for* w *as well as a set of type-shifting rules, which are themselves terms of* $\Lambda\mathsf{Ty}_n$*.*

Formulas of the Lambek calculus (Lambek, 1958) are inductively defined from a set of atomic formulas np (noun phrase), n (common noun), s (sentence) and pp (prepositional phrase)[2]. A formula in the Lambek calculus is

[2]The set of atomic formulas used is slightly more detailed than this, see (Moot, 2010a) for details. However, for the discussion in the current paper, this set of atomic formulas will suffice.

$$\frac{A/B : f^{U\to T} \quad B : x^U}{A : (fx)^T}\ /E$$

$$\frac{\begin{array}{c}\ldots[B : x^U]\\ \vdots \\ A : t^T\end{array}}{A/B : \lambda x^U t}\ /I$$

$$\frac{A/B : f^{\Pi\alpha.U[\alpha]\to T} \quad B : x^{U[V]}}{A : (f\{V\}x)^T}\ /E^*$$

Figure 1: Proof rules and corresponding lambda term operations

- an atomic formula
- if A and B are formulas, then A/B (pronounced "A over B", it looks for a B formula to its left to produce an A) and $B\backslash A$ (pronounced "B under "A", it looks for a B formula to its left to produce an A) are formulas.[3]

Figure 1 shows the proof rules for the Lambek calculus; only the rules for '$/$' are shown, the rules for '$\backslash$' are left-to-right symmetric for those of '$/$'.

The elimination rule for '$/$', labeled '$/E$' states that if we have a proof with conclusion A/B which is assigned term f (of type $U \to T$) and a proof with conclusion B which is assigned term x (of type U), then we can combine these two proofs to form a proof of A which is assigned lambda-term $(f\ x)$.

The introduction rule, labeled '$/I$', states that if we have a proof of A with lambda-term t of some type U, which we have derived while using a hypothesis B, which is assigned a variable x of type U and which is the rightmost undischarged hypotheses of this proof, then without this B, we can derive A/B of type $\lambda x.t$.

This correspondence between natural deduction proofs and lambda-terms is the well-know Curry-

[3]We will not consider the product formulas $A \bullet B$ in this paper.

Howard correspondence (it is not an isomorphism for the Lambek calculus) .

The '$/E^*$' rule is a variation of the '$/E$' rule which permits a limited type of polymorphism: it allows a function to specify its argument only partially. The lambda-term assigned to "le" (the) in Figure 2 is an instance of this type scheme.

5 Lexical Semantics

Our semantic approach is generated in the tradition of lexical semantics called the Generative Lexicon (Pustejovsky, 1995), especially in its type-logical interpretation of (Bassac et al., 2010)

In our type hierarchy, have two specific types of spatial arguments *regions* and *paths*. Two functions *source* and *destination* convert a path p to its source region and its destination region[4]. We also assume a spatial variable *here* which denotes the position and orientation of the spatial reference point (which does not necessarily correspond to the position of the narrator; in this sense it is closer to a spatial equivalent of the Reichenbachian "reference time" than it is to the constant "now": in a DRS it is most naturally implemented as a succession of values as is the reference time t). Both the position and orientation are necessary to understand a discourse like the following.

(10) a. My new apartment is awesome.
b. The entrance hall is spacious.
c. To the left, there is the living room.

In this discourse, we can make sense of the expression "to the left" only because we make a kind of "virtual visit" with up/down, forward/backward and left/right well-defined.

The distinction between regions and paths is rather standard (Jackendoff, 1983). It is motivated by selectional restrictions on verbs: some verbs, such as "stay + PP" are only grammatical when the PP is a preposition denoting a region argument, whereas other verbs, such as "pass + PP" can only occur with a number of PPs, all of which denote a path. This distinction is muddled slightly by the possibility to coerce a region r into a path. As is well known, some prepositions, such as "vers" (towards), can — at least in their spatial uses — only denote paths.

For our semantic analysis, we interpret all motion verbs as being relations between one or more entities and a *path*. This argument can be left implicit (ie. when we say "John ran"). Verbs specify lexically which of their arguments follow this path (subject, object or both, see (Nam, 1995)).

"Le chemin" with type assignment $np - \iota x^{immobile\_object}.chemin(x)$ does not combine with "descend" which requires a person as its argument, as indicated by its lexical entry which is of the general form $np\backslash s - \lambda y^{person} \ldots$ (to focus the discussion on the coercion mechanism, we give only a schematic entry at this point, the complete entry is shown in Figure 2). Both "chemin" and "descend" permit lexically anchored type coercions: "chemin" has a lexical lambda term g which coerces it in such a way that "le chemin" obtains type assignment $np - \iota x^{path}.chemin(x)$ (again simplified for readability), whereas "descend" has a lexical lambda term h which coerces its lexical semantics to $np\backslash s - \lambda y^{path} \ldots$. Note that applying both coercions makes "le chemin descend" a correctly typed term, with "le chemin" being a term of type $path$ and "descend" a term of type $path \rightarrow t$.[5]

Full details on the lexical type assignments can be found in Figure 2.[6]

Some comments about the lexical semantics. $\oplus$ is the DRS merge operation. Motion verbs are analysed by the "neutral" predicate *travel* (in the style of (Miller and Johnson-Laird, 1976)) which takes an event e, a moving entity x and a path p as its arguments. It is true if the traveler x follows path p during event e. The functions *source* and *destination* are functions from paths to their source and destination regions, whereas *height* is a

[4]We are aware that there are many ways to refer to places in the *middle* of the paths as well. However, we assume that this is done by a relation *middle*(p, x), where p is a path and r a region, rather than by functions. Note that *source*, *middle* and *destination* give us a way of encoding the difference between Initial, Median and Final verbs in the terminology of (Asher and Sablayrolles, 1995).

[5]We abstract away from the notions of rigid and flexible coercions of (Bassac et al., 2010; Mery, 2011).

[6]The type assignments have been slightly simplified. The type of a DRS should be $s \rightarrow s \rightarrow t$ (ie. "state updates") instead of t (ie. simple truth values), see (Muskens, 1994; Muskens, 1996) for details.

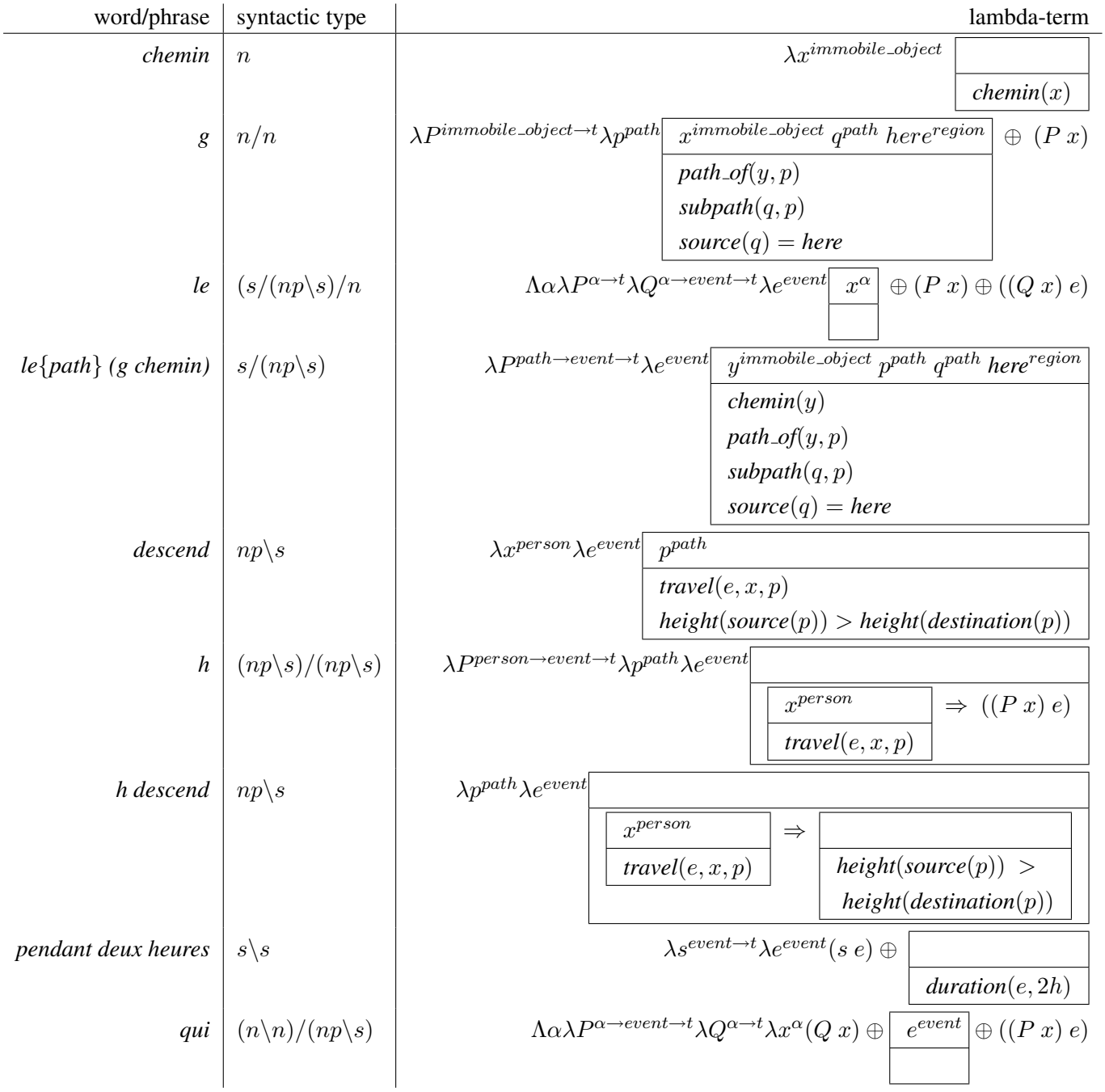

Figure 2: Lexical types and lambda terms for “le chemin descend pendant deux heures”, together with the coercions g and h.

function from regions to their vertical coordinate. Taken together, the semantic entry for "descend" states therefore that given an argument which is a person x and an argument which is an event e, the DRS will verify that there is a path p such that x follows p and that the height at the start of this path p is greater than his height at the end of it.

The coercion h, which is lexically specified for "descend" but not necessarily for other motion verbs which are less naturally interpreted as describing properties of the corresponding path, allows us to change the interpretation of "descend" into stating: if a person follows the path p, then he descends, as shown by the beta-reduced lexical entry for "h descend" in the figure.[7]

Note that the entry for "h descend" does not commit us to concluding that anyone actually takes the path. This must be deduced separately.

The lexical entry for the determiner "le" is assigned a polymorphic type. It selects a noun to produce a quantifier; however since the type of this term can vary it is of type $\alpha \rightarrow t$, where α is determined by the noun which serves as the argument for the determiner, eg. $\alpha = immobile\_object$ in the case of "chemin" and $\alpha = path$ after the coercion. Note that a simple application of the $/E^*$ rule shown in Figure 1 suffices in either case, instantiating the variable α according to the type of the noun.

The lexical entry for "chemin" is rather simple and indicates only that x is an entity of type immobile object for which *chemin*(x) holds (this is a rather standard — but also rather trivial — semantics in the tradition of Montague). What is more interesting is that the lexical coercion g allows us to change the type of the argument from an immobile object x to a path p, while asserting that this path corresponds to the immobile object as indicated by the predicate *path_of* and selecting a sub-path q of p going forward from *here*, which may or may not go to the end of the path p. Note that having both x (the immobile physical object aspect to the path) and p (its path aspect) as referents in the universe of the DRT is necessary to account for modifiers of both aspects of the path, as in "a brick road to Pau" (see also sentence (9)), as well to allow anaphoric references to both aspects of the path, as in the following example.

(11) a. The street was completed in 1825 (...)
b. It runs from the Regent's residence at Carlton House (...) to All Souls Church.

The use of the variable *here*, which has both a place and an orientation, has the pleasant consequence of there being no incoherence between saying "le chemin monte" and "le chemin descend" at exactly the same place but with just the orientation reversed.

The prepositional phrase "pendant deux heures" (for two hours) is given a rather simple Davidsonian analysis: it simply states that the duration of the event corresponding to the sentence it modifies is two hours.

It should be noted, however, that, in spite of the fact that the analysis we have presented assigns them similar meanings, the following two sentences should have a rather different interpretation.

(12) The path descends abruptly/slowly.

(13) John descends abruptly/slowly.

In sentence (13) John is going down the path and he is doing so abruptly or slowly which is neutral with respect to the slope of John's path: eg. John can descend a steep slope slowly.

By contrast, sentence (12) *does* allow us to infer that the slope of the path either suddenly becomes rather steep ("descends abruptly") or has a rather level downwards slope for the contextually relevant stretch of it.

This difference is partly explained by the lack of an agent in sentence (12): if there is no consciousness guiding the movement then the abruptness can only come from external factors. Another way of interpreting these facts is to see sentence (12) as talking about a "generic" traveler taking this path, who therefore also has a "default" means of transport which can be deduced from the context. This default means of transport is necessary

[7] To save space, the entry for "h descend" is slightly simplified. The DRS condition *travel*(e, x, p) has been removed from the right-hand side of the implication. However, since *travel*(e, x, p) already occurs on the left-hand side, this makes no difference for the interpretation of the DRS

for the correct interpretation of "for two hours" as well. The discussion of adverbs has stayed rather informal and we admit that we only have sketched some possible solutions. However, given the difficulty of the semantics of adverbs, a more detailed and formal treatment of adverbs would be a much larger research project.

6 Computing Discourse Representation Structures

So far, we have only treated simple sentences without much of the surrounding context. The real test for this analysis is how it interacts with the constraints on interpretation posed by its surrounding context.

(14) (...) nous descendons, pendant un quart d'heure, la vallée de l'Esera.
we descend, for a quarter of an hour, the Esera valley.

(15) La lune, qui éclaire notre marche, nous fait découvrir sur la droite un sentier qui serpente.
The moon, which lightens our steps, allows us to discover a winding path on our right.

(16) Il nous conduit sur un petit plateau, au milieu de sapins, au-dessus et à quelque distance du torrent de Ramun.
It leads us to a small plateau, surrounded by firs, at some distance of and above the Ramun torrent.

Here, "Il" (it) in sentence (16) refers to "un sentier qui serpente" (a winding path) for the previous sentence, so a correct analysis requires resolution of the anaphor before coercion in order to give a correct analysis.

As is well known, the rhetorical relations as used by SDRT (Asher and Lascarides, 2003), provide a set of important constraints on the possible interpretations on discourse. In the example above, we have the relations *Background*(14,15) and *Narration*(15,16).

A second example illustrates the importance of rhetorical structure on the interpretation (as well as the difficulty of automatically obtaining such a structure).

(17) Nous partimes pour Barèges à 8 heures du matin par une fort jolie route qui nous conduisit à Lourdes.
We left (PS) for Barèges at 8 in the morning, taking a very pretty road which led (PS) us to Lourdes.

(18) (...) qui va en se resserrant jusqu'à Pierrefite, où les routes de Lux et de Cauterets séparent.
(...) which goes shrinking along the way, up to Pierrefite, where the roads to Lux and to Cauterets split.

(19) Celle de Lux entre dans une gorge qui vous mène au fond d'un précipice et traverse le gave de Pau.
The one to Lux enters a gorge which leads you to the bottom of a precipice and traverses the Gave de Pau.

(20) (...) Après une longue marche, l'on arrive à Barèges à 6 heures du soir.
(...) After a long walk, we arrive in Barèges at 6 in the evening.

Here, sentence (17) introduces the destination and therefore the whole spatio-temporal extension route. The following will therefore constitute an Elaboration relation between this sentence and the sequence of (18)-(20). It is (at first sight) difficult to decide on the discourse relation of Sentence (19): it would certainly be possible to have a later phrase beginning with "Celle de Cauterets" (the road leading to Cauterets) and a number of the following sentences (omitted here for space reasons) give further background information about the road to Lux. However, at sentence (20), it suddenly becomes evident that the author has been describing the road while following it.

7 Conclusions and Future Work

We have given a treatment of "virtual movement" in a type-logical grammar. Our account merges two successful extensions of "standard" Montague-style semantics — DRT and the Generative Lexicon — into a single, coherent type-theoretic framework.

Though we have provided some solutions which

can easily be implemented — Emeric Kien has provided a prototype implementation of the coercion mechanism, which we plan to extend — some of the problems we have touch upon in the article, such as anaphora resolution and determining the appropriate discourse relations between segments of text, will require a more significant effort.

References

Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for French. In Anne Abeillé, editor, *Treebanks: Building and Using Parsed Corpora*, chapter 10, pages 165–187. Kluwer, Dordrecht.

Nicolas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambrdige University Press.

Nicholas Asher and Pierre Sablayrolles. 1995. A typology of discourse semantics for motion verbs and spatial PPs in French. *Journal of Semantics*, 12(2):163–209.

Nicolas Asher, Philippe Muller, and Mauro Gaio. 2008. Spatial entities are temporal entities too: The case of motion verbs. In *LREC 2008 Workshop on Methodologies and Resources for Processing Spatial Language*, Marrakech, May 2008.

Nicholas Asher. 2011. *Lexical Meaning in Context. A Web of Words*. Cambridge University Press.

Christian Bassac, Bruno Mery, and Christian Retoré. 2010. Towards a type-theoretical account of lexical semantics. *Journal of Logic, Language and Information*, 19(2):229–245.

Johan Bos, Stephen Clark, Mark Steedman, James R. Curran, and Julia Hockenmaier. 2004. Wide-coverage semantic representation from a CCG parser. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-2004)*, pages 1240–1246, Geneva, Switzerland.

Jean-Yves Girard, Yves Lafont, and Paul Taylor. 1988. *Proofs and Types*. Number 7 in Cambridge Tracts in Theoretical Computer Science. Cambridge University Press.

Ray Jackendoff. 1983. *Semantics and Cognition*. MIT Press.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. Kluwer Academic Publishers, Dordrecht.

Joachim Lambek. 1958. The mathematics of sentence structure. *American mathematical monthly*, pages 154–170.

Pierre Loustau. 2008. *Interprétation automatique d'itinéraires dans des récits de voyages D'une information géographique du syntagme à une information géographique du discours*. Ph.D. thesis, Université de Pau.

Bruno Mery. 2011. *Modélisation de la sémantique lexicale dans le cadre de la théorie des types*. Ph.D. thesis, Université de Bordeaux.

George A. Miller and Philip Johnson-Laird. 1976. *Language and Perception*. Harvard University Press.

Richard Moot and Christian Retoré. 2011. *The Logic of Categorial Grammars*. Lecture Notes in Artificial Intelligence. Springer. to appear.

Richard Moot. 2010a. Semi-automated extraction of a wide-coverage type-logical grammar for French. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*, Montreal.

Richard Moot. 2010b. Wide-coverage French syntax and semantics using Grail. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*, Montreal.

Reinhard Muskens. 1994. Categorial Grammar and Discourse Representation Theory. In *Proceedings of COLING 94*, pages 508–514, Kyoto.

Reinhard Muskens. 1996. Combining Montague Semantics and Discourse Representation. *Linguistics and Philosophy*, 19:143–186.

Seungho Nam. 1995. *The Semantics of Locative Prepositional Phrases in English*. Ph.D. thesis, UCLA.

James Pustejovsky. 1995. *The generative lexicon*. M.I.T. Press.

Leonard Talmy. 1999. Fictive motion in language and "ception". In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, pages 211–276. MIT Press.

Johan van Benthem. 1987. Categorial grammar and lambda calculus. In D. Skordev, editor, *Mathematical Logic and Its Applications*, pages 39–60. Plenum Press, New York.