Traitement automatique sur corpus de récits de voyages pyrénéens : Une analyse syntaxique, sémantique et temporelle

Anaïs Lefeuvre[1,2,3], Richard Moot[1,2], Christian Retoré[1,2], Noémie-Fleur Sandillon-Rezer[1,2]
Université de Bordeaux[1], LaBRI-CNRS[2] & INRIA [3]
anais.lefeuvre@labri.fr, moot@labri.fr, retore@labri.fr, nfsr@labri.fr

RÉSUMÉ

Cet article présente notre utilisation de la théorie des types dans laquelle nous nous situons pour l'analyse syntaxique, sémantique et pour la construction du lexique. Notre outil, Grail permet de traiter le discours automatiquement à partir du texte brut et nous le testons sur un corpus de récit de voyages pyrénéens, Itipy. Nous expliquons donc notre usage des grammaires catégorielles et plus particulièrement du calcul de Lambek et la correspondance entre ces catégories et le λ-calcul simplement typé dans le cadre de la DRT. Une flexibilité du typage doit être autorisée dans certains cas et bloquée dans d'autres. Quelques phénomènes linguistiques participant à une forme de glissement de sens provocant des conflits de types sont présentés. Nous expliquons ensuite nos motivations d'ordre pragmatique à utiliser un système à sortes et types variables en sémantique lexicale puis notre traitement compositionnel du temps des évènements inspiré du *Binary Tense* de (Verkuyl, 2008).

ABSTRACT

Processing of a Pyrenees travel novels corpus : a syntactical, semantical and temporal analysis.

In this article, we present a type theoretical framework which we apply to the syntactic analysis and the computation of DRS semantics. Our tool, Grail, is used for the automatic treatment of French text and we use a Pyrenées travel novels corpus, Itipy, as a test case. We explain our use of categorial grammars and specifically the Lambek calculus and its connection to the simply typed λ-calculus in connection with DRT. Flexible typing has to be allowed in some cases and forbidden in others. Some linguistic phenomena presenting some kind of meaning shifts inducing typing conflicts will be introduced. We then present our motivations in the pragmatic field to use a system with sorts and variable types in lexical semantics and then we present how we process events temporality, in the light of Verkuyl's *Binary Tense*(Verkuyl, 2008)

MOTS-CLÉS : compositionalité, interface syntaxe-sémantique, interface sémantique-pragmatique, grammaire catégorielle, théorie des types, récit de voyage.

KEYWORDS: compositionality, syntax-semantics interface, semantics-pragmatics interface, categorial grammar, type theory, travel novel.

Ce travail de recherche a reçu un soutien financier d' INRIA et du Conseil Régional d'Aquitaine dans le cadre du projet Itipy

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 2: TALN, pages 43–56,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Cet article décrit les étapes qui composent notre analyse du discours, en partant du texte brut pour en produire une représentation sémantique dans le cadre de la *Discourse Representation Theory* (DRT) (Kamp et Reyle, 1993). Une chaine complète de traitement est proposée et testée sur le corpus Itipy. Ce corpus de récits de voyage pyrénéens a été rassemblé par la médiathèque de Pau pour mettre en valeur le fond patrimonial de récits de voyage dans sa région.

Une analyse du discours impose de fait une interaction entre la sémantique des unités de langue dont on doit interpréter le sens en discours et la prise en compte de la dimension pragmatique de ce qui est dit (Busquets *et al.*, 2001). Certains phénomènes sémantiques restent difficiles à traiter, certains cas de glissement de sens montrent qu'une flexibilité dans le typage doit être permise, alors que dans les cas les plus courants le typage doit être rigide pour éviter une représentation inappropriée. Nous donnons quelques exemples et proposons, afin d'améliorer les résultats de notre chaîne de traitement, de traiter ces phénomènes par l'affinement des λ-termes du lexique dans le cadre d'un systèmes à sortes et types variables, dans le λ-calcul d'ordre supérieur.

Dans le cadre du projet Itipy, nous nous sommes intéressés à la dimension temporelle des évènements dans le discours, ce qui nous a amené à interroger la compositionnalité de cet aspect. En nous appuyant sur les travaux de (Verkuyl, 2003), nous voulons introduire dans le traitement un système de test puis un lexique propre aux locutions adverbiales de temps. Nous avons introduit les termes en λ-calcul qui permettent de déterminer la valeur temporelle d'un évènement. On compose le terme propre à la phrase que l'on applique aux termes propres des adverbes, puis aux opérateurs perfectif ou imperfectif, postérieur ou simultané, et enfin présent ou passé en fonction de la morphologie du verbe conjugué. Nous expliquerons plus en détails ce système et traiterons deux exemples de notre corpus illustrant ce système.

Nous détaillerons d'abord notre corpus et nos objectifs applicatifs quant à celui-ci, nous présenterons les étapes de traitement du discours, commençant par l'acquisition de la grammaire du français sur corpus annoté, puis l'analyse syntaxique dans le cadre des grammaires catégorielles. Nous expliquerons plus amplement l'interface syntaxe-sémantique dans la théorie des types logiques permettant la construction de nos représentations sémantiques en λ-DRT. Nous présenterons brièvement le système de types variables et notre traitement des phénomènes discursifs en jeu dans l'interaction sémantique-pragmatique, ainsi que le traitement temporel des évènements.

2 Le corpus

Notre corpus de 576 334 mots est une collection de 11 oeuvres classées par la médiathèque de Pau comme récits de voyages pyrénéens du XIXème et début XXème siècle. Concernant les données textuelles de notre corpus, le genre du récit de voyage, implique de fait une hétérogénéité interne reconnue. Certains spécialistes désignent par ailleurs le récit de voyage comme un "genre fragmenté" (Magri-Mourgues, 2009), dans lequel on trouve une myriade de procédés narratifs incluant "le récit métonymique" , "le récit synecdochique", "le récit métaphorique", "le récit de voyage et de découverte du réel", etc. Ajoutons à ceci que le corpus Itipy est constitué de récits écrits par des géologues, des topographes, ou encore des romanciers.

Malgré la diversité des formes de discours qui composent le corpus, sa spécificité réside dans

le récit de l'itinéraire, seul point commun entre tous les textes. La structure narrative du récit de voyage observe une alternance entre la description de l'itinéraire emprunté et d'autres informations telles que des observations sur le relief, le caractère des personnages rencontrés ou encore des considérations introspectives du narrateur sur des domaines variés.

Notre traitement du discours se détache totalement des données du genres, ou encore de la nature des thèmes abordés. Nous intervenons alors sur l'analyse profonde syntaxique et sémantique, mais aussi pragmatique et discursive du discours. En partant d'une automatisation de l'analyse syntaxique, la représentation sémantique est construite automatiquement elle aussi utilisant pour ressource un lexique sémantique saisi préalablement à la main. Afin de produire des représentations du discours satisfaisantes, nous travaillons sur un affinement des λ-termes contenus dans ce lexique.

3 Analyse syntaxique

Grail est un analyseur pour grammaires catégorielles dans la tradition de Lambek (Lambek, 1958) et leurs extensions multimodales (Moortgat, 1997). Dans des travaux récents (Moot, 2010a,b), Grail a été étendu pour l'analyse du français à large couverture. La figure 1 montre les composants de la chaîne de traitement pour le français : il y a un *part-of-speech tagger*, un *supertagger* (Clark et Curran, 2004) pour limiter le nombre de formules que l'analyseur doit traiter et un lexique sémantique qui associe à chaque mot un λ-terme correspondant à son type. Dans le lexique, on utilise des λ-termes produisant des DRS après substitution et normalisation (Muskens, 1994).

3.1 Acquisition du lexique syntaxique et apprentissage des modèles d'entropie maximale

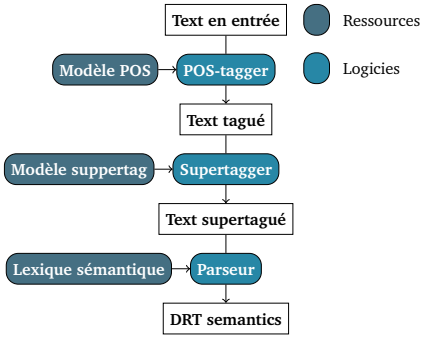


FIGURE 1 – Schéma des ressources et outils de la chaîne de traitement

Le *French Treebank* (Abeillé *et al.*, 2003) a été transformé, en partie automatiquement, en dérivations pour grammaires catégorielles. La complexité résidait dans le fait que les arbres du *French Treebank* sont plats, avec un nombre de fils maximal par noeud non fixé, alors que les arbres de dérivations doivent être binaires. La figure 2 montre une sous-partie d'un arbre du corpus. Des techniques standard pour extraire des grammaires catégorielles à partir d'arbres d'annotation (Buszkowski et Penn, 1990) nécessitent des arbres binaires avec, pour chaque paire de frères, une indication pour la tête, et une pour l'argument : si la tête est à gauche, les formules correspondant au deux frères sont A/B et B, si la tête est à droite ce sont B et $B\backslash A$, où la formule B dépend du syntagme du noeud dans l'arbre d'annotation (eg. NP correspond à la for-

mule atomique np et INF correspond à la formule complexe $np\backslash s_{inf}$). Les arbres du corpus sont alors binarisés et des heuristiques déterminent la tête d'un syntagme, faisant un effort pour rester le plus fidèle possible aux analyses habituelles en grammaire catégorielles ; typiquement le verbe est la tête d'une phrase, le déterminant la tête d'un groupe nominal (pour rester cohérent avec une possible analyse sémantique ultérieure), etc. On calcule ainsi récursivement les formules, de la racine aux feuilles, où les feuilles donnent les formules pour les mots du lexique.

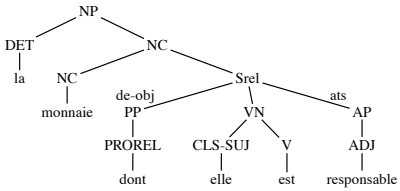


FIGURE 2 – Exemple d'un arbre planaire du *French Treebank*, avant traitement

Une méthode utilisée pour extraire une grammaire catégorielle du *French Treebank*, et donc un lexique représentatif de celui-ci, est l'utilisation d'un transducteur d'arbre. Le principe du transducteur d'arbre, explicité dans (Comon *et al.*, 1997), est de prendre en entrée un arbre tel que montré dans 2 et et restituer un nouvel arbre en sortie. Pour ce travail, nous nous sommes limités à la partie *AB* d'une grammaire de Lambek, qui correspond aux mécanismes les plus basiques d'une langue naturelle. Les grammaires AB sont une référence en inférence grammaticale (Buszkowski et Penn, 1990) et c'est ce qui a motivé notre choix premier. La sortie de ce transducteur [1] est donc une forêt d'arbres de dérivation d'une grammaire AB. A partir de celle-ci, nous pouvons soit extraire un lexique (voir figure 3), qui contient les mots des phrases analysées, les différents types trouvés et leur occurrence, soit une grammaire (voir figure 5) d'arbres qui contient à la fois les règles que nous considérons comme correctes (étant donné que ce sont celles qui apparaissent dans nos arbres de dérivation) et des probabilités sur ces règles. Le lexique peut servir à entrainer le Supertagger.

$$8374 : la \quad \rightarrow \quad 7996 : np/n, \quad 94 : (n\backslash n)/n, \quad 57 : (s\backslash s)/n, \quad 43 : (s/s)/n, ...$$

FIGURE 3 – Extrait du lexique. Le mot "la" est utilisé 8374 fois dans la partie analysée du corpus. La catégorie la plus fréquente correspond à celle d'un déterminant qui attend un nom commun à sa droite pour créer un groupe nominal. Les trois types suivant correspondent à des modificateurs, comme "**La** semaine dernière …" en début de phrase.

Il ne reste ensuite qu'à extraire le λ-terme correspondant à l'arbre. Cette méthode présente cependant des limitations, et certains phénomènes, tels que les traces, ou les ellipses ne se traitent pas avec des grammaires AB. Sans que cela limite le nombre de phrases analysées, les types donnés aux mots s'en trouvent complexifiés.

Ces traces, qui sont des dépendances non-bornées, ne sont pas indiquées dans le corpus et ont été ajoutées à la main dans une phase de nettoyage post-traitement. De plus, une phase de correction manuelle est nécessaire pour éliminer certaines différences entre les analyses choisies pour le corpus et les analyses habituellement utilisées pour les grammaires catégorielles (voir (Moot, 2010a)). Les modifications sont entre autre l'ajout de trace (il y a plus de 500 occurences de "que/qu"), la restructuration des groupes verbaux (l'argument du noyau verbal devint argument uniquement du participe passé lorsqu'il y a lieu).

1. Le transducteur que nous avons mis en place est détaillé dans (Sandillon-Rezer et Moot, 2011).

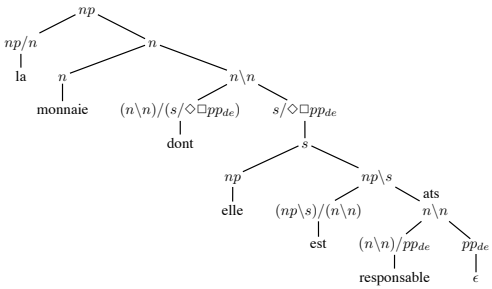


FIGURE 4 – résultat de l'extraction

Elles permettent de réduire le nombre de formules du lexique, qui passent de 5240 à 918, et donc le nombre d'analyses possibles. La figure 4 montre le résultat de l'extraction : les feuilles s'ajoutent au lexique.

Pour l'entrainement des taggers, 11.196 phrases (334.525 mots) sont utilisées et 1.244 phrases (36.504 mots) pour l'évaluation. Les modèles atteignent une précision de 98,4% pour le tagger et de 90,5% pour le supertagger (Moot, 2010a).

s	$\rightarrow np\ np\backslash s$	21,56%
s	$\rightarrow s/s\ s$	7,18%
...		

FIGURE 5 – Extrait de la grammaire. On trouve les règles binaires ayant généré la forêt d'arbres, groupées par racine (ici s) et le pourcentage d'utilisation de celles-ci.

3.2 Analyseur

Comme indiqué dans la figure 1, après une *tokenization* simple, les mots d'une phrase d'entrée sont d'abord étiquetés par le tagger. Au niveau des étiquettes utilisées lors de la *tokenization*, on utilise celles de Treetagger, car elles donnent des informations sur le temps des verbes, ce qui va nous servir pour le traitement du temps, sans pour autant utiliser Treetagger pour l'analyse.

Le supertagger sert surtout comme filtre à l'analyseur propre : en utilisant un facteur β ($\beta \leq 1$), réglable par l'utilisateur, et en se fixant sur le type le plus probable pour un mot, il va sélectionner les types dont les probabilités sont au delà de $\beta.proba\_max$. Ainsi, plus β est petit, plus les types ayant une faible probabilité sont proposés à l'analyseur. Ceci permet d'ajouter plus de formules pour des mots considérés comme difficiles (c'est-à-dire pour lesquels le modèle a relativement peu confiance en son premier choix), mais garde une seule formule pour les mots considérés comme faciles.

Ensuite, l'analyseur se charge des combinaisons des formules selon les règles du calcul de Lambek multimodal, utilisant les formules lexicales les plus probables prioritairement. Pour la grammaire à large couverture, Grail ne garde que la première analyse trouvée. Et cette analyse, qui correspond à un lambda terme simplement typé, va servir pour calculer la sémantique. Remarquons que ce calcul est très simple : on utilise simplement la β-réduction (avec peu ou pas de duplication de termes) pour obtenir une forme normale qui correspond à une formule ou, dans notre cas, à une DRS.

4 Représentation sémantique en λ-DRT

A l'instar de Boxer, développé par (Curran *et al.*, 2007), Grail associe à l'analyse syntaxique, une représentation sémantique dans le style de la DRT. La DRT est une théorie proposant de représenter la sémantique d'un discours grâce à un modèle présenté comme une boîte (*Discourse Representation Structure*) dans laquelle on trouve d'un coté le domaine, composé des référents du discours, quantifiés implicitement par un existentiel, et de l'autre les conditions d'interprétation sémantique de ce modèle. Cette théorie permet de construire ces structures mettant en évidence la valeur sémantique, d'un point de vue logique, des éléments porteurs de sens d'un énoncé au sein d'un contexte. Les descriptions sémantiques en λ-DRT de chaque item constituent un lexique écrit à la main préalablement et qui sert de ressource pour la chaine de traitement.

Abordons maintenant la correspondance entre catégories dans le calcul de Lambek et propriétés calculatoires et logiques des λ-termes. L'isomorphisme Curry-Howard montre que les dérivations en grammaires catégorielles sont des sous-ensembles des dérivations de la logique intuitionniste. Autrement dit, à chaque analyse catégorielle de phrase correspond un λ-terme normal. Les catégories syntaxiques dans le style du calcul de Lambek (*s, sn/n*) permettent donc d'associer une lecture sémantique exprimée par le λ-terme simplement typé fourni par le lexique, dans le style de la DRT, et correspondant à chaque mot taggé. Au sein de la structure syntaxique, on associe à sa catégorie le terme en λ-DRT associé puis on β-réduit l'expression. Les DRS créées se fusionnent les unes avec les autres par l'opération de "*merge*" et permettent d'interpréter les phénomènes de cohérence du discours, comme par exemple la résolution des anaphores pronominales. En dérivant l'analyse sémantique de l'analyse syntaxique, on conserve la bonne formation de la représentation de l'expression correspondant exactement à la catégorie.

4.1 De l'analyse syntaxique à la forme logique

En premier lieu, il convient de définir ce qu'est un type selon la sémantique de Montague. L'analyse syntaxique de type *s* présentée précédemment, est un λ-terme dont les variables libres correspondent aux mots. Le lexique fournit des λ-termes du même type sémantique. En les substituant, et en réduisant le terme obtenu, on obtient un terme normal de type **t** : c'est une formule logique, la représentation sémantique, et dans notre cas la λ-DRS. Néanmoins il faut au minimum partager le type **e**, les individus (aussi appelés entités) [2], en diverses sortes pour que le calcul de la sémantique bloque à juste titre lorsque le type d'un argument ne correspond pas au type attendu par la fonction.

Pour notre système de test, nous avons implémenté un petit jeu d'adverbes temporels comme *dans x heures*, *en x minutes*, *pendant x jours*, etc. Nous souhaitons à terme intégrer les outils tels que (Bittar, 2009), (Parent *et al.*, 2008). Nous précisons que χ (chronos) est la fonction prenant un évènement et renvoyant un intervalle temporel. Prenons deux exemples issus du corpus et le lexique grammatical, syntaxique puis sémantique correspondant à chacun des items : Les formules ici présentées sont des formules de logique partielles dont nous détaillerons la construction dans la section 4.3.

(1) *Le 31, nous sommes partis à six heures du matin.*

2. Nous utilisons les variables e dans les exemples suivant pour désigner des évènements et non des personnes.

item lexical	POS	catégorie	λ-terme
Le 31	ADV	s/s	$\lambda s\lambda e.(s\ e) \wedge \chi e \subseteq jour(31)$
nous	PRO :PER	sn	$nous$
sommes	VER :pres	$(sn\backslash s)/(sn\backslash s_{ppart})$	PRES (PERF)
partis	VER :pper	$sn\backslash s_{ppart}$	$\lambda x\lambda e.partir(e,x)$
à six heures du matin	ADV	$s\backslash s$	$\lambda s\lambda e.(s\ e) \wedge \chi e \circ 6\ :00$

(2) *Dans dix minutes, j'aurai quitté Nohant.*

item lexical	POS	catégorie	λ-terme
Dans dix minutes	ADV	s/s	$\lambda s\lambda e.(s\ e) \wedge distance_{min}(\chi e, n) = 10$
j	PRO :PER	sn	je
aurai	VER :fut	$(sn\backslash s)/(sn\backslash s_{ppart})$	PRES(POST)(PERF)
quitté	VER :pper	$(sn\backslash s_{ppart})/sn$	$\lambda y\lambda x\lambda e.quitter(e,x,y)$
Nohant	NAM	sn	$Nohant$

Les λ-termes manipulant les intervalles χe sont inspirés des préconisations de (Verkuyl, 2003) qui propose un traitement des adverbes. Par exemple pour *dans dix minutes*, on utilise une fonction qui donne la distance la plus courte en minutes entre le moment d'énonciation n et l'intervalle repère de l'évènement (R chez Reichenbach), tout comme pour PRES, POST et PERF, nous reviendrons plus en détail sur les termes associés à ces opérateurs dans la section 4.3.

Pour simplifier la présentation du lexique, *je*, *nous* et *Nohant* sont présentées comme des constantes. *partir* est un prédicat à deux arguments, respectivement un évènement nommé e et un agent, x. On remarque la correspondance entre le sn attendu à gauche dans la catégorie syntaxique et la présence d'un argument agentif dans la représentation sémantique par le λ-terme. *nous* appliqué au terme *partir* une fois β-réduit ne contiendra qu'un seul λe nécessaire à la manipulation temporelle de l'évènement. *quitter* est un prédicat à trois arguments, respectivement un évènement nommé e, un agent, x et une source (argument spatial) y, de la même manière on remarque les deux sn attendu d'abord à droite pour la source puis à gauche pour l'agent.

Le typage du lexique permet de vérifier de manière stricte la bonne formation de la représentation mais ne permet pas d'interpréter certaines expressions du discours parfois plus souples de ce point de vue, c'est pourquoi nous proposons une solution dans la suite des travaux de Pustejovsky (Pustejovsky, 1995) et d'autres issus de la même tradition.

4.2 Un raffinement de la sémantique lexicale : sortes et types variables

Dans un travail initié par Bassac, Rétoré et Mery (Bassac *et al.*, 2010) dédié à la partie sémantique lexicale de l'analyse, et assez proche en surface à la réinterprétation de Markus Egg (Egg, 2002), une nouvelle organisation du lexique a été proposée afin de traiter, dans la sémantique formelle compositionnelle, les glissements de sens, l'accès aux différentes facettes du sens, et la possible coprédication. On peut observer ces phénomènes sur les exemples suivants :

(3) ** La chaise aboie.* Ce genre de composition impossible est rejetée par un système de types plus riche : *aboie* a pour argument un *chien* à la rigueur un *humain* mais jamais un *meuble*.

(4) *Ce livre est volumineux mais intéressant.* Coprédication correcte entre les deux facettes de *livre* : contenu informationnel et objet physique.

(5) ** Grenoble a battu Dax et prévoit d'acquérir pour 474 500 euros d'œuvres d'art.* Coprédication impossible entre le club de rugby et la municipalité.

Pour ces phénomènes courant dans notre corpus, nous avons conçu une structure de lexique et un algorithme qui permettent de calculer les représentations sémantiques de telles phrases, de rendre compte des coprédications correctes (4) d'échouer lorsqu'elles ne le sont pas (5) et de quantifier correctement. Les types de base des λ-termes, qui sont les sortes d'une logique multisorte, servent à éviter les compositions impossibles comme 3. En cas de conflits de type, il est parfois licite de relaxer le typage, ou d'accéder à l'une des facette du sens d'un mot : une ville peut être considérée comme sa municipalité ou son club de rugby, un livre est à la fois un objet physique et informationnel, etc. Pour ce faire, le lexique associe à chaque mot le λ-terme usuel ainsi que plusieurs λ-termes permettant de changer le type du mot, et certaines transformations, comme celle de la ville en club sont déclarées irréversibles, ce qui permet de prédire l'impossibilité de certaines coprédications. Ce genre de phénomènes nécessite, en particulier pour la conjonction à l'œuvre dans les phénomènes de coprédication, des opérations uniformes sur les types, et c'est pour cela que nous nous sommes placés dans le système F (Girard, 1971). Ce formalisme nous permet de manipuler plus finement les types, de quantifier sur eux, d'utiliser la coercion afin de résoudre les cas de coprédication par exemple. Cette organisation du lexique permet de traiter bien des phénomènes, de sémantique lexicale mais aussi de sémantique compositionnelle.

Les conflits se présentent alors sous la forme $(\lambda x^A.u)w^W$: un terme de type A est attendu par la fonction $(\lambda x^A.u)$ mais l'argument fourni est de type W. Pour résoudre ces conflits, lorsque cela conduit à des interprétations licites, on utilise les λ-termes optionnels du lexique qui donnent au mot le type correspondant au sens adéquat, de l'une des deux manières suivantes :

Transformation rigide correspondant aux transformations irréversibles incompatibles avec les autres transformations. Le lexique fournit, pour un mot de u ou pour un mot de w un λ-terme g de type $W \to A$: le terme se résout en $(\lambda x^A.u)(gw)^A$.

Transformation flexible correspondant aux transformations compatibles avec les autres transformations. Les diverses occurrences de x^A dans u sont utilisées avec des types différents $A_1, \ldots, A_n$: on peut utiliser, si le lexique en fournit, des termes différents de types $g_i : W \to A_i$ pour chaque occurrence de x et remplacer comme le veut la β-réduction chaque occurrence de x par $(g_i(w)) : A_i$.

En sémantique lexicale, ce modèle nous permet même de traiter de constructions assez subtiles, comme le *voyageur fictif* où un *chemin* introduit un *voyageur* qui le suivrait.

(6) *Pendant deux heures le chemin descend.(On notera que c'est le circonstanciel qui oblige à considérer un voyageur fictif sinon cela ne serait pas nécessaire.)*

Là encore, c'est le conflit de type : $\left(P^{humain \to \mathbf{t}}\left(u^{chemin}\right)\right)$ $\quad humain \neq chemin$ qui déclenche l'utilisation du λ-terme optionnel associé à chemin : celui-ci transforme la route en un événement, dont l'agent un voyageur fictif (voir (Moot *et al.*, 2011)).

Cette proposition aborde aussi deux questions classiques de sémantique formelle, dont le traitement s'intègrent dans la même proposition, la quantification généralisée et les pluriels : il s'agit ici de construire les formules logiques associées à certains énoncés et non de déterminer leurs conditions de vérité dans tel ou tel modèle.

Voici tout d'abord un exemple de pluriels correspondant à un quantificateur généralisé :

(7) En effet, on est ici voisin de Toulouse ; comme le caractère, le type est nouveau. Les jeunes filles ont des figures fines, régulières, d'une coupe nette, d'une expression vive et gaie. Elles sont petites, elles ont la démarche légère, des yeux brillants, la prestesse d'un oiseau.

Ici, définir les filles de la région comme ayant des "figures régulières", étant "petites" ou encore ayant la "prestesse d'un oiseau" est une comparaison sous entendue à une fille prototypique de

cet ensemble de filles. On conceptualise facilement une idée de la taille comme étant normale pour un spécimen du type "fille". Ainsi il nous faut un opérateur pouvant sélectionner toutes les propriétés telle que la taille d'un type particulier afin de pouvoir l'associer au spécimen de ce type, quelque soit le type concerné. En premier lieu, dans ce système nous rappelons qu'il n'existe qu'un quantificateur peu importe la classe d'objet sur laquelle on quantifie, ce qui permet de quantifier sur tous les ordres. Au lieu d'avoir une constante $\forall_\alpha$ de type $(\alpha \rightarrow \mathbf{t}) \rightarrow \mathbf{t}$ pour chaque type α sur lesquels on voudrait quantifier, le quantificateur est donc $\forall$ de type $(\alpha \rightarrow \mathbf{t}) \rightarrow \mathbf{t}$ pour tout type α et il sera ensuite spécialisé au type désiré. La quantification généralisée "la plupart des" ou "les" est prise en charge par une constante $\measuredangle$, à rapprocher du $\tau x.A$ d'Hilbert : étant donné un type α, notre constante $\measuredangle$ renvoie le spécimen du type α — pour plus de détails voir (Retoré, 2012).

Notre deuxième exemple concerne toujours les pluriels, mais lorsqu'on prédique une propriété d'un ensemble d'individus, ce qui suscite plusieurs interprétations :

(8) Edgar et son guide descendaient toujours ensemble !... Enfin, le groupe allait se briser sur une saillie de roc effrayante, quand Vincent se précipita avec intrépidité au-devant d'eux, enfonçant par un coup désespéré sa hache tout entière dans la neige...

Cet exemple du corpus permet d'observer un phénomène bien connu, où un ensemble d'entités agit collectivement ou au contraire réunit des entités agissant individuellement. Ainsi on peut comprendre dans "le groupe allait se briser sur une saillie de roc effrayante" que la chute sépare les deux individus qui composait le groupe, dans ce cas c'est le groupe qui se brise ou encore que chacun d'entre eux subit les dommages de l'accident, et alors ce sont les individus appartenant au groupe qui sont brisés. Suivant l'interprétation choisie, la valeur de vérité sera vraie pour l'un et fausse pour l'autre. Notre système à sortes et types variables permet de gérer cette difficulté et les deux interprétations grâce à la constante de distributivité présentée ici.

La constante * : $\lambda P^{\alpha \rightarrow t} \lambda Q^{\alpha \rightarrow t} \forall x^{\alpha}.Q(x) \Rightarrow P(x)$ qui peut être spécialisée à n'importe quel type α permet une distributivité de la propriété sur les membres de l'ensemble. Les détails formels concernant cette constante et d'autres gérant la coercion et la distributivité stricte sont décrits dans (Moot et Retoré, 2011)

Cette technique permet de respecter le principe selon lequel la syntaxe guide la composition sémantique de l'énoncé. Ici, le raffinement lexical permet de filtrer les interprétations impossibles et résoudre ces conflits lorsque le lexique le permet : ainsi on rejette les interprétations erronées sans rejeter d'interprétations obtenues pas des glissements de sens ou l'accès à des facettes du sens. Ces mécanismes ont été ajoutés à la sémantique de Grail en λ-DRT — cette dernière a été étendue au λ-calcul du second ordre. L'implantation — sur de petits lexiques — a été réalisée par Emeric Kien (Kien, 2010) et elle est actuellement poursuivie par Samira Kherfellah.

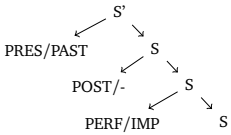
4.3 Le traitement temporel

La temporalité en DRT est traditionnellement (Partee, 1984; Kamp et Reyle, 1993) traitée par des relations entre constantes inspirées des constantes de Reichenbach. Ces modélisations du temps des évènements sont interprétées par des intervalles et des points. Nous nous sommes demandés dans quelle mesure on pouvait retrouver le principe de compositionnalité que Reichenbach ne réussit pas à conserver intégralement. Rappelons simplement les unités utilisées sur l'axe du temps : le point d'énonciation, l'intervalle de l'évènement et le point de repère en cours (respecti-

vement *S*,*E* et *R*) (Reichenbach, 1948). Tout d'abord dans (Verkuyl, 2003), puis dans (Verkuyl, 2008), l'auteur propose une approche complètement compositionnelle, des combinaisons de trois choix entre deux λ-termes complémentaires dont on donne l'abbréviation :
– PAST / PRESENT
– SYNCHRONOUS / POSTERIOR
– PERFECTIF / IMPERFECTIVE
Chaque terme combiné aux deux autres permet de traiter les huit temps verbaux néerlandais à l'indicatif ainsi que les huit temps verbaux anglais correspondant. Le français dispose de six temps supplémentaires dans le mode indicatif dont le passé simple, le passé antérieur et les formes surcomposées du passé et du futur. Toutes ces formes sont traitées par le système. Seule la représentation de la sémantique temporelle de l'évènement est mise en évidence par chacun des termes, n'est pas traité la classe aspectuelle par ces combinaisons. Ils nous permet de procéder à l'analyse temporelle des évènements de manière compositionnelle et conforme à notre analyse en λ-DRT. Nous avons construit la grammaire décrite dans (Verkuyl, 2003) puis nous l'avons intégré dans notre système.

Définissons notre langage inspiré des recommandations de Verkuyl pour traiter de la temporalité des évènements, ce langage étant un sous langage des relations de Allen, il se définit comme ceci : $\langle$ n,i_1,..., i_x, χ, $\circ$, $<$,=, $\subseteq$, $\supset\subset$ $\rangle$
n est une constante
i_1,..., i_x sont des variables
Les lambda termes des opérateurs sont définis et s'appliquent comme suit :

$$\begin{array}{rll} \text{PRES} & =_{def} & \lambda\phi\exists i(\phi\ i)\wedge(\chi i)\circ n \\ \text{PAST} & =_{def} & \lambda\phi\exists i(\phi\ i)\wedge(\chi i)< n \\ \text{POST} & =_{def} & \lambda\phi\lambda i\exists j(\phi\ j)\wedge(\chi i)<(\chi j) \\ \text{PERF} & =_{def} & \lambda\phi\lambda j\exists k(\phi\ k)\wedge(\chi k)<(\chi j) \\ \text{IMP} & =_{def} & \lambda\phi\lambda j\exists k(\phi\ k)\wedge(\chi j)\subseteq(\chi k) \end{array}$$



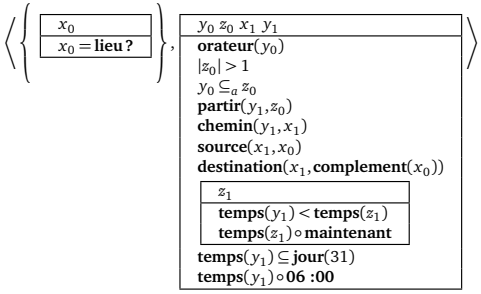
On applique le λ-terme du prédicat évènementiel muni de ses arguments au λ-terme de l'opérateur comme pour n'importe quelle autre unité du lexique. Les adverbes temporels quant à eux doivent intervenir soit avant tout opérateur et donc au plus près du noyau prédicatif, soit entre PERF/IMP et POST/ –. Concernant le typage de ces opérateurs, il est définit sur les valeurs **i**, le type accordé à l'indice chez Verkuyl, qui dans notre cas est le type de l'intervalle [3], et sur la valeur **t** pour la valeur de vérité. Le type des opérateurs sera donc : $\mathbf{i}\rightarrow\mathbf{t}\rightarrow\mathbf{t}$ auquel on applique le prédicat évènementiel avec ses arguments, lui même de type : $\mathbf{i}\rightarrow\mathbf{t}$.
Intuitivement, le n de Verkuyl joue un rôle similaire au S de Reichenbach, le i semble être un R pour le PAST, mais c'est plutôt le j pour le PERF. Une formule de ce langage est donc formée de ces variables, en relation ou non avec la constante, ce qui permet de construire les termes pour les adverbes temporels et pour la temporalité attachée au verbe, on construira la combinaison de plusieurs opérateurs tel que PAST(POST), PRES(POST(PERF)), etc.

Le modèle dans lequel on veut interpréter ce langage est l'ensemble des intervalles tel que l'a définit (Allen, 1983), on donne ici la traduction des relations :

$$\begin{array}{l|l|l|l} \|<\| = \{<,\text{m}\} & \|\subseteq\| = \{\text{s,d,f,=}\} & \|=\| = \{=\} & \text{On remarque que la relation } y\circ n \\ \|>\| = \{>,\text{mi}\} & \|\supseteq\| = \{\text{si,di,fi,=}\} & \|\supset\subset\| = \{\text{m}\} & \text{veut dire } y \text{ di } n. \end{array}$$

3. C'est pourquoi nous avons la fonction χ, soit **temps**( ) dans les DRS prenant un évènement et renvoyant un intervalle dans le lexique, on ne s'étendra pas sur les adverbes qui peuvent affecter deux indices au lieu d'un seul.

Penchons nous désormais sur notre exemple 1 : *Le 31, nous sommes partis à six heures du matin.*



(9)

Pour ce premier exemple, on observe plusieurs choses, tout d'abord, du point de vue spatial, tous les verbes de déplacement doivent être envisagés comme accompagnés d'un chemin défini par une *source* et une *destination*, tous deux liés à la temporalité de l'évènement. La source étant le lieu dans lequel se situe le voyageur au début de l'évènement et la destination, le lieu occupé à la fin de l'évènement de déplacement.

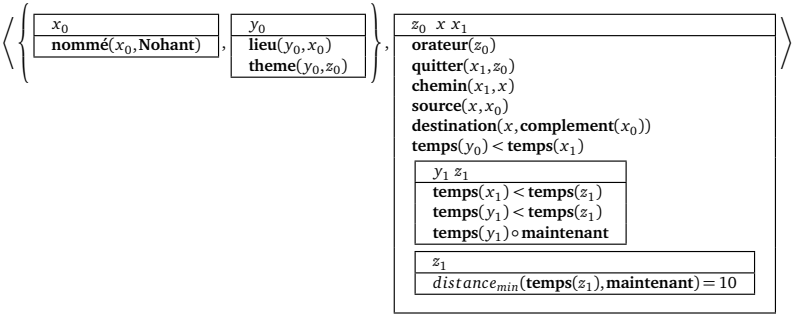
Il y a par ailleurs quelques présuppositions qu'il faut éclairer. On peut inférer de *quitter x* qu'avant de quitter, le voyageur est dans le lieu désigné par x. Par ailleurs, *complément (x)*, argument de *destination* dans le cas de quitter, tout comme dans le cas de partir dans l'exemple ci-après, sera donc la région de l'espace dans laquelle on se trouve une fois avoir quitté x ou être parti de x. On ne peut pas en déduire une destination à proprement dit mais pour le moins on peut désigner par *complément (x)* l'extérieur de x. Nous notons une seconde remarque concernant les présuppositions spatiales profondes nécessaires à l'expression de *quitter* et *partir* selon laquelle, elles résistent aux épreuves de la modalisation induite par *désirer* et de la négation :

(10) Après avoir visité aussi le Vignemale et le Pic du Midi de Bigorre, je désirai ne point quitter les Pyrénées sans avoir fait du moins un effort en faveur de l'ascension de la Maladetta.

Cet exemple ne peut être interprété que comme l'énoncé de quelqu'un situé dans les Pyrénées au moment de l'évènement dénoté par *désirer*.

Il nous faut aussi interroger la nature même de *partir*, considère-t-on la finalité de partir comme le fait de n'être plus là ou bien dans le fait d'être dans le mouvement du départ ? On réfère ici à la distinction entre *accomplishment* et *achievement* de Vendler (Vendler, 1967), typiquement si on dit *je suis presque parti* alors on a affaire à un *achievement* car l'événement sera réalisé selon la condition qu'on ne soit plus dans le lieu d'origine. Si on dit *je partais lorsqu'elle s'est adressée à moi* ici l'essence de l'évènement peut durer et donc être considéré comme un *accomplishment*.

Au sujet d'*à six heures du matin*, la relation ∘ choisie entre l'intervalle *06 :00* et l'intervalle y_1 est plutôt faible et on ne peut véritablement rien inférer d'autre au sujet des deux entités en relation. Concernant *Le 31* il a fallut décider si la valeur accordée à ce syntagme adverbial devait être donnée comme étant dans le futur ou dans le passé. En effet, ici l'interprétation dépend entièrement du contexte. On imagine sans peine un exemple utilisant aussi cet item tel que : "*Le 31 je serai aux Galapagos*" qui impliquerait de fait une interprétation dans le futur. Que faire alors des expressions de l'habitude telles que "*Le vendredi je mange du poisson*" ? Est-ce une résolution, ou une habitude de longue date ? A ce sujet nous pensons intégrer à terme un module traitant de la SDRT dans le système afin d'obtenir les informations propres aux relations discursives pouvant résoudre ce problème. Les expressions telles que *le vendredi*, *le 31* méritent que l'on s'interroge davantage sur leur sémantique en contexte.
Regardons de plus près notre second exemple 2 : *Dans dix minutes, j'aurai quitté Nohant.*



(11)

Pour calculer correctement et respecter l'accessibilité des variables dans la représentation, nous avons fait le choix d'imbriquer les DRS propres à PRES, POST et à PERF décrits plus tôt. Plus exactement si l'on décompose le calcul, *(PRES(POST(PERF (quitter e, je, Nohant))))*, on obtient le terme suivant :
$\exists y_1 \exists z_1 \exists x_1 [quitter(x_1, je, Nohant)] \wedge (\chi x_1) < (\chi z_1) \wedge (\chi y_1) < (\chi z_1) \wedge (\chi y_1) \circ n)]$
interprété comme :
– PRES : par rapport au moment d'énonciation n, il existe un repère y_1 qui est en relation d'$\circ$
– POST : par rapport à ce repère, il existe un intervalle postérieur y_1
– PERF : l' intervalle y_1 est lui même postérieur à la fin de l'intervalle de l'évènement x_1
Dans la figure 6, on donne une représentation graphique possible.

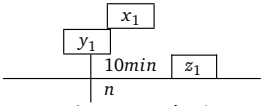


FIGURE 6 – Interprétation possible des variables temporelles pour l'exemple traité.

La transitivité des relations ne nous donne que peu d'informations sur la localisation de l'évènement *quitter*, par exemple x_1 n'a finalement qu'une seule contrainte par rapport à n (maintenant) si ce n'est que tout comme y_1, il doit se situer avant z_1 qui lui même est en relation *di* avec n pour Allen soit $\circ$ ici.

Ce système permet d'exprimer la temporalité des évènements portée par le verbe conjugué ainsi que les modifications que peuvent apporter les adverbes. L'un des atouts de ce système réside dans le fait que l'on ne dit pas plus de choses dans la formule que n'est dit dans le discours. La représentation sémantique de la temporalité de chaque évènement est compositionnelle et propose un contenu précis et approprié au traitement des relations entre les évènements, étape suivante de nos travaux. Les premiers tests opérés semblent prometteurs mais ce travail nécessite un enrichissement du lexique pour les adverbes et des tests à plus grande échelle.

5 Conclusion

Dans cet article nous avons montré les différentes étapes de notre traitement automatique du discours, consistant en l'analyse syntaxique puis en la dérivation sémantique en λ-DRT. L'interface

syntaxe-sémantique dans le cadre de la théorie des types est une base solide permettant de respecter la compositionnalité du sens tout en s'appuyant de l'organisation syntaxique du discours. Le système F quant à lui est approprié pour traiter les phénomènes rencontrés dans le discours et l'interface sémantique-pragmatique justifie un raffinement du lexique par ce système. Pour de futurs travaux, nous envisageons d'enrichir davantage le lexique afin de couvrir plus largement le discours et les phénomènes de glissement de sens ainsi que les modifications temporels. Dans le cadre du projet Itipy, il est nécessaire de développer davantage l'ordonnancement temporel des évènements dans le récit en déterminant les relations appropriées et choisir les composantes temporelles qui doivent être mises en relation. Dans le cadre du genre de discours étudié et de l'objet que nous cherchons à extraire, la classe aspectuelle de chaque prédicat muni de ses arguments doit être déterminée en fonction de son rapport à l'espace afin de relier correctement les évènements qu'ils dénotent. Plus concrètement, nous envisageons en premier lieu de développer une composante permettant la résolution d'anaphores, première étape indispensable à la suite de nos travaux.

Références

Abeillé, A., Clément, L. et Toussenel, F. (2003). Building a treebank for french. *In Treebanks*. Kluwer.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *In Communications of the ACM*, numéro 26(11), pages 832–843.

Bassac, C., Mery, B. et Retoré, C. (2010). Towards a Type-Theoretical Account of Lexical Semantics. *Journal of Logic Language and Information*, 19(2):229–245.

Bittar, A. (2009). Annotation of events and temporal expressions in french texts. *In The Third Linguistic Annotation Workshop (LAW III) Singapore*.

Busquets, J., Vieu, L. et Asher, N. (2001). La SDRT : Une approche de la cohérence du discours dans la tradition de la sémantique dynamique. *Verbum*, XXIII(1):73–101.

Buszkowski, W. et Penn, G. (1990). Categorial grammars determined from linguistic data by unification. *Studia Logica*, 49:431–454.

Clark, S. et Curran, J. R. (2004). Parsing the WSJ using CCG and log-linear models. *In Proceedings of the 42nd annual meeting of the ACL*, pages 104–111, Barcelona.

Comon, H., Dauchet, M., Jacquemard, F., Lugiez, D., Tison, S. et Tommasi, M. (1997). Tree automata techniques and applications.

Curran, J., Clark, S. et Bos, J. (2007). Linguistically motivated large-scale nlp with c&c and boxer. *In Proceedings of the 45th Annual Meeting of the ACL*, pages 33–36, Prague.

Egg, M. (2002). Semantic contructuction for reinterpretation phenomena. *Linguistics*, 40(3): 579–609.

Girard, J.-Y. (1971). Une extension de l'interprétation de Gödel à l'analyse et son application : l'élimination des coupures dans l'analyse et la théorie des types. *In* Fenstad, J. E., éditeur : *Proceedings of the SLS*, volume 63 de *Studies in Logic and the Foundations of Mathematics*, pages 63–92, Amsterdam. North Holland.

Kamp, H. et Reyle, U. (1993). *From Discourse to Logic*. D. Reidel, Dordrecht.

KIEN, E. (2010). Du sens des mots à l'analyse automatique d'une phrase. Mémoire de stage d'initiation à la recherche, ENS-Cachan & INRIA Bordeaux.

LAMBEK, J. (1958). The mathematics of sentence structure. *American Mathematical Monthly*, 65:154–170.

MAGRI-MOURGUES, V. (2009). *Le voyage à pas comptés. Pour une poétique du récit de voyage au XIXème siècle*. Numéro 9 de Lettres numériques. Honoré Champion.

MOORTGAT, M. (1997). Categorial type logics. *In* van BENTHEM, J. et ter MEULEN, A., éditeurs : *Handbook of Logic and Language*, chapitre 2, pages 93–177. Elsevier/MIT Press.

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

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

MOOT, R., PRÉVOT, L. et RETORÉ, C. (2011). Un calcul de termes typés pour la pragmatique lexicale. *In Traitement Automatique du Langage Naturel, TALN 2011*, pages 161–166, Montpellier.

MOOT, R. et RETORÉ, C. (2011). Second order lambda calculus for meaning assembly : on the logical syntax of plurals. *In Coconat 2011*.

MUSKENS, R. (1994). Categorial Grammar and Discourse Representation Theory. *In Proceedings of COLING 94*, pages 508–514, Kyoto.

PARENT, C., GAGNON, M. et MULLER, P. (2008). Annotation d'expressions temporelles et d'évènements en français. *In Traitement automatique des langues naturelles*.

PARTEE, B. H. (1984). Nominal and temporal anaphora. *Linguistics and Philosophy*, 7:243–286.

PUSTEJOVSKY, J. (1995). *The Generative Lexicon*. MIT Press.

REICHENBACH, H. (1948). *Elements of Symbolic Logic*. The Mac millan Company.

RETORÉ, C. (2012). Variable types for meaning assembly : a logical syntax for generic noun phrases introduced by "most". *Recherches linguistiques de Vincennes*, 41:1–18.

SANDILLON-REZER, N.-F. et MOOT, R. (2011). Using tree tranducers for grammatical inference. *Proceedings of Logical Aspects of Computational Linguistics 2011*, pages 233–250.

VENDLER, Z. (1967). *Linguistics in philosophy*. Cornell University Press.

VERKUYL (2003). On the compositionnality of tense : Merging reichenbach and prior. Utrecht University.

VERKUYL, H. (2008). *Binary Tense*. CSLI Publications.