Modélisation de la Sémantique Lexicale dans le cadre de la théorie des types

Bruno Mery

Bruno Mery
Université Bordeaux 1
LaBRI, Laboratoire Bordelais de Recherche en Informatique
350, cours de la Libération
F-33405 Talence Cédex

Thèse Doctorale de Recherche
Spécialité Informatique

Modélisation de la sémantique lexicale dans le cadre de la théorie des types

Mots clés : Lexique génératif, Ontologies, Sémantique compositionnelle, λ-calcul du second ordre.

Soutenue le Mardi 5 Juillet 2011

Membres du Jury :

Directeurs :	Christian Bassac et Christian Retoré
Rapporteurs :	Nicholas Asher et Violaine Prince
Examinateurs :	Guy Mélançon et Sylvain Pogodalla
Invitée :	Alexandra Arapinis

Bruno Mery
Université Bordeaux 1
LaBRI, Laboratoire Bordelais de Recherche en Informatique
350, cours de la Libération
F-33405 Talence Cédex

Philosophiæ Doctorate Thesis
In Computer Science

Modelling lexical semantics in a type-theoretic framework

Keywords : Generative lexicon, Ontologies, compositional semantics, second-order λ-calculus.

Defended on Tuesday, July 5th 2011

Members of the Jury

Supervisors : Christian Bassac and Christian Retoré
Reviewers : Nicholas Asher and Violaine Prince
Examiners : Guy Mélançon et Sylvain Pogodalla
Guest : Alexandra Arapinis

In Memoriam
Paul Mery
1952-2009

Il n'y a, en réalité, ni vérité ni erreur, ni oui ni non, ni autre distinction quelconque, tout étant un, jusqu'aux contraires.
Il n'y a que des aspects divers, lesquels dépendent du point de vue.

— Zhuang Zi.

La science et son objet diffèrent de l'opinion et de son objet, en ce que la science est universelle et procède par des propositions nécessaires, et que le nécessaire ne peut pas être autrement qu'il n'est.

— Aristote.

Wise words are like arrows flung at your forehead. What do you do ? Why, you duck of course.

— Steven Erikson.

Résumé
Le présent manuscrit constitue la partie écrite du travail de thèse réalisé par Bruno Mery sous la direction de Christian Bassac et Christian Retoré entre 2006 et 2011, portant sur le sujet « Modélisation de la sémantique lexicale dans la théorie des types ». Il s'agit d'une thèse d'informatique s'inscrivant dans le domaine du traitement automatique des langues, et visant à apporter un cadre formel pour la prise en compte, lors de l'analyse sémantique de la phrase, d'informations apportées par chacun des mots.

Après avoir situé le sujet, cette thèse examine les nombreux travaux l'ayant précédée et s'inscrit dans la tradition du lexique génératif. Elle présente des exemples de phénomènes à traiter, et donne une proposition de système de calcul fondée sur la logique du second ordre. Elle examine ensuite la validité de cette proposition par rapport aux exemples et aux autres approches déjà formalisées, et relate une implémentation de ce système. Enfin, elle propose une brève discussion des sujets restant en suspens.

Abstract
This paper is part of the thesis by Bruno Mery advised by Christian Bassac and Christian Retore in the years 2006-2011, on the topic "Modelling lexical semantics in a type-theoretic framework". It is a doctoral thesis in computer science, in the area of natural language processing, aiming to bring forth a formal framework that takes into account, in the parsing of the semantics of a sentence, of lexical data.

After a discussion of the topic, this thesis reviews the many works perceding it and adopts the tradition of the generative lexicon. It presents samples of data to account for, and gives a proposal for a calculus system based upon a second-order logic. It afterwards reviews the validity of this proposal, coming back to the data samples and the other formal approaches, and gives an implementation of that system. At last, it engages in a short discussion of the remaining questions.

Remerciements
La rédaction du présent manuscrit n'a pas été un travail solitaire. Si tous les manquements de cette thèse sont de mon unique responsabilité, chacune des avancées, chaque fragment d'idée, et surtout la volonté nécessaire à les accomplir, sont des impulsions d'autres personnes, qu'elles en soient ou non conscientes.

C'est pourquoi je tiens à remercier, tout d'abord, mes encadrants ; Christian Bassac, qui a été d'une aide indispensable tout au long de la rédaction, et Christian Retoré, sans qui le calcul des termes à l'ordre supérieur n'existerait pas ; Émeric Kien, pour son travail réalisé ; toutes les personnes qui m'ont conseillé, et notamment Michele Abrusci, Nicholas Asher, Patrick Henry, Alexandre Köller, Richard Moot, Sylvain Salvati, tous les membres de l'équipe Signes et tous les chercheurs que j'oublie ; les nombreux auteurs qui m'ont inspiré, James Pustejovsky au premier chef ; mes amis, dont Julien Arnould, Nada Ayad, Pascal Barthoumieux, Nicolas Bellino, Thomas Bieber, Cyril Boisnier, Stéphane Boucley, Pierre Bourreau, Guillaume Capdupuy, Clément Charpentier, Mélanie Claudot, Florian Coquart, Patrick Decollas, Cécile, Cyril et Élias Dumange, Béatrice Dumora, Sana Hakim, Michel Fournier, Estelle Inacio, Amandine Jambert, Kristian Kocher, Jory Lafaye, Jean-David Laffitte, Isabelle Le Maistre, Anaïs Lefuvre, Steven LeDelliou, Yannick Mary, Samuel Méril, Stéphanie Moreaud, Frank Morféa, Guillaume Pascual, Gaël Prado, Fanny Sallen, Noémie-Fleur Sandillon-Rezer, François Seimandi, Guillaume Vidal, Natalia Vinogradova, et tous ceux que j'ai oubliés, qui m'ont soutenu et sans cesse encouragé. J'aimerais également remercier mes professeurs qui, au long de ma scolarité puis de mes études, m'ont permis d'avancer vers ce sujet de thèse, même si je ne l'ai probablement pas traité suffisamment bien pour eux : Vincent Beamonte, Anne Dicky, Irène Durand, Géraud Sénizergues, Robert Strandh, Alexandre Zvonkine et de très nombreux autres ; pour des raisons personnelles, je souhaite aussi remercier M. Ferrière, M. Nivet et Mme Tanguy. Sur une note plus détendue, je tiens aussi à remercier l'ensemble des équipes du CROUS d'Aquitaine, dont le café et thé sont pour beaucoup dans ma productivité. Je remercie chaleureusement mes chats, nommées Reality et Entity dans un moment difficile de progrès dans les présents travaux, pour leur présence et le réconfort qu'elles m'apportent.

Enfin, je remercie ma famille pour leur soutien inconditionnel, et en premier lieu ma mère, Christiane Mery, qui m'a conduit plus que tout autre à m'investir au-delà des possibilités de ma simple volonté dans un travail dont je ne voyais plus l'aboutissement.

Ce manuscrit est dédié à la mémoire de mon père, Paul Mery, sans qui je serais bien loin de la personne que je suis aujourd'hui.

Table des matières

Chapitre 1

Introduction

En préambule

La langue est, encore aujourd'hui, considérée comme l'apanage de l'Homme, la plus grande réalisation de l'espèce ; par elle, on écrit l'histoire et invente les histoires, on engrange le savoir aussi bien que les faux-semblants. L'étude de la langue peut sembler sacrilège, car il s'agit d'essayer de comprendre comment les êtres humains se comprennent, et de transcrire de façon claire ce processus afin qu'une machine puisse s'en saisir. Mais l'Informatique a bien ce but, car elle est la science du traitement de l'information ; l'analyse du langage est une étape essentielle de son développement.

J'ai pour intime conviction que l'humanité se doit de comprendre le monde qui l'entoure et de se comprendre elle-même. Savoir pourquoi nous pouvons nous comprendre, et comment nous sommes à même d'employer un même mot dans de multiples sens selon le contexte et rester capables de produire un discours cohérent, produire une formalisation, c'est-à-dire une machine, qui puisse faire de même, tel est mon but dans le présent manuscrit.

Les travaux de cette thèse de doctorat ne répondent pas avec satisfaction à toutes les questions que je me suis posé à l'origine ; le manuscrit en lui-même est probablement trop court pour une thèse de cette prétention. Cependant, ils forment un tout cohérent et constituent une avancée, un pas de plus vers une sémantique formelle précise et complète.

Ces travaux ont été encadrés par deux directeurs de thèse, effectués au sein d'un laboratoire et d'un organisme de recherche, avec de nombreuses équipes et de nombreux collaborateurs. Durant leur réalisation, j'ai bénéficié de conseils de chercheurs et d'experts et de l'appui de très nombreuses personnes et institutions ; c'est pourquoi, dans tout le reste de ce manuscrit, j'emploierai le « nous » collectif pour désigner les auteurs de toutes les réflexions ayant conduit à cette réalisation finale.

Ces travaux sont également inachevés, et devront être poursuivis, par moi-même ou par d'autres, durant les longues années que nécessitera la pleine réalisation de leurs ambitions… Mon espoir est qu'ils n'aient pas été vains.

Bruno Mery, Talence. 2011.

Historique

Les réflexions que nous allons présenter ici portent sur l'étude formelle de la langue, et en particulier du lexique. Il peut être utile de commencer par nous replacer dans un contexte historique riche en développements. Une grande partie des références développées ici sont issues de [Marcel Cori, 2002], qui étudie l'historique des changements de dénominations des disciplines concernées dans l'histoire récente.

Nul ne sait à quand remontent les premières réflexions sur le lexique, mais nous disposons de traces datant d'avant notre ère tendant à montrer que cette origine se perd dans l'antiquité. Tandis qu'en Inde, Panini donnait une description formelle de la grammaire des Veda, en Mésopotamie, les rois Achéménides gravaient leurs tablettes de lois en trois langues, donnant le premier exemple de traduction simultanée et proposant un lexique trilingue. Leur contemporain Grec, Aristote, est le fondateur des principes ayant donné lieu aux théories que nous prenons pour point de départ.

Mais il fallut longtemps avant de considérer la langue comme un formalisme. Ferdinand de Saussure, à la fin du 19^{eme} siècle (publié postmortem dans [Bally and Sechehaye, 1916]) décrit pour la première fois le langage en tant que *système de signes* ; c'est le début de la linguistique structuraliste, qui se développera par la suite en de multipes courants.

Pour ce qui est des modèles informatiques, c'est Petr Smirnov-Trojanski qui déposera, en 1933, une première demande de brevet à Moscou pour une *machine à traduire*, demande rejetée car les brevets, dans la Russie Soviétique de l'époque, n'étaient accordés que si le demandeur peut faire la preuve que l'invention ne pouvait être utilisée massivement. En 1935, Georges Artsrouni propose, lui aussi, une machine à traduire utilisant une bande perforée pour encoder un lexique bilingue, sans aucun calculateur ; sa machine obtiendra un grand prix à l'Exposition Universelle de Paris, en 1937.

Parallèlement, les travaux d'Alan Turing (1936) et de Von Neumann (1945) contribuent à la création de l'informatique en tant que science.

C'est la volonté de traiter le langage comme les calculateurs traitent l'arithmétique, et le problème particulier de la traduction, qui est à l'origine des premiers travaux mêlant linguistique et informatique.

En 1947, Wiener aborde, dans une correspondance avec Weaver, la question de la traduction mécanique. Dans les années 1948-49, Shannon et Weaver contribuent à l'élaboration de la théorie de l'information, et c'est en 1949 que Warren Weaver rédige un Memorandum intitulé *Translation*, qu'il transmet à une trentaine de connaissances ; c'est le point de départ de recherches à grande échelle sur la traduction automatique.

Parallèlement, en 1950, Turing présente son Test qui permettrait de déterminer si une machine peut être déclarée « intelligente », ce test fait appel aux compétences langagières de l'individu.

Des séries de publications sur la traduction automatique suivent. En 1950, Erwan Reifler publie ainsi *Studies in Mechanical Translation*. Bar-Hillel préside le colloque du MIT intitulé *Conference on Mechanical Translation* en 1952. *Mechanical Translation*, une revue savante, est fondée en 1954 par Victor Yngve, et, la même année, IBM fait la première démonstration de traduction par un ordinateur à New York ; suite à cette démonstration, la National Science Foundation et la CIA décident de financer de nouvelles recherches, et Panov convainc l'URSS de faire de même.

Des groupes de recherches apparaissent entre 1955 et 1962 au Japon, en Tchéquoslovaquie, en Chine, en France, en Italie, au Mexique, en Belgique…

En 1956, la conférence d'été de Darthmouth College signe la naissance de l'Intelligence Artificielle en tant que discipline.

La linguistique n'est pas en reste puisque Noam Chomsky publie en 1957 *Syntactic Structures*, qui reste l'ouvrage de référence de la grammaire transformationelle-générative.

En France, les recherches s'organisent en 1959 avec la création de l'ATALA (Association pour l'étude et le développement de la Traduction Automatique et de la Linguistique Appliquée) à l'initiative de Delavenay, de l'UNESCO ; du CETA (Centre d'Études pour la Traduction Automatique) à l'institut Blaise Pascal par le CNRS et la défense nationale ; du Centre de Linguistique Quantitative par Jean Favard, de l'institut Henri Poincarré.

Cependant, les recherches se poursuivent et n'aboutissent pas aussi rapidement qu'envisagé à leur origine. C'est Yeoshua Bar-Hillel qui jette un premier pavé dans la mare en 1960, avec son rapport intitulé *A Demonstration of the Nonfeasibility of Fully Automatic High Quality Translation*. Il s'appuie alors sur des exemples qui ne sont pas traités par la traduction automatique, comme *The box was in the pen*, dans lequel *pen* dispose de deux sens différents (en français : *crayon*, *parc*) suivant le contexte. Bar-Hillel argue que, pour un contexte donné (comme *Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.*), le locuteur peut sans mal donner le sens approprié, tandis qu'une machine en sera incapable à moins de disposer d'un lexique comportant un savoir encyclopédique, proposition à laquelle Bar-Hillel répond « *This is surely utterly chimerical and hardly deserves any further discussion.* » (c'est pourtant de ce genre d'exemples que naîtra le domaine de la sémantique lexicale).

En 1960, l'ATALA initie la revue *Traduction Automatique*. En 1962 est fondée l'AMTCL (*Association for Machine Translation and Computational Linguistics*).

Parallèlement, Joseph Weizembaum, du MIT, publie en 1964 le programme ELIZA, première simulation d'un dialogue homme-machine fondée sur la psychologie humaine.

Mais la notion de traduction automatique tombe peu à peu en disgrâce. En 1965, l'ATALA abandonne cette dénomination pour devenir l'*Association pour le Traitement Automatique du Langage* et sa revue devient *TA information, revue internationale du Traitement Automatique du Langage*. Aux États-Unis, en 1966, l'ALPAC (*Automatic Language Processing Advisory Committee*) publie un rapport qui conduit à l'arrêt des subventions pour la recherche en *Machine Translation*, ce qui aura des répercussions dans le monde entier. Le même rapport préconise cependant le maintien du financement pour la recherche dans la « traduction aidée par les machines », ce qui contribue à imposer la discipline qui sera désormais connue sous le nom de *Computational Linguistics*.

La recherche se tourne désormais vers la formalisation, le but étant de représenter les connaissances, de modéliser l'analyse de la langue. Ainsi seront développés, dans les années 1970 et 1980, de nouvelles formes de logiques (logique floue, logique modale…), les réseaux sémantiques de Quillian (1966), la *frame semantics* de Marvin Minsky (1974), les *scripts* de Roger Schank (1977), les graphes conceptuels de John Sowa (1984)…

En France, en 1984, André Lentin propose l'adoption de l'expression « Linguistique Computationnelle » pour qualifier la discipline résultant des multiples transformations des années passées.

Depuis les années 1990, la recherche s'est concentrée sur la linguistique de corpus. Les changements de dénominations continuent en France, où, en 1993, l'ATALA change le nom de sa revue en TAL (*Traitement Automatique des Langues*) et organise en 1994 la première conférence du TALN (*Traitement Automatique du Langage Naturel*).

Depuis, les recherches continuent, et l'objectif de la compréhension de la langue par l'ordinateur semble aussi loin qu'aux origines.

Sujet d'études

Les applications du traitement formel de l'information se sont étendues bien au-delà des aspirations de ses précurseurs. Recherche, industrie et commerce ont contribué à faire des dispositifs de traitement des objets de la vie quotidienne plus que de coûteux dispositifs expérimentaux, et de l'Informatique, au lieu d'un domaine restreint des mathématiques, une science *sui generis* aux foisonnants sujets de recherche et d'applications. Cependant, de nombreuses questions fondamentales, posées depuis les premiers travaux du domaine, demeurent toujours sans réponse.

Ce manuscrit s'inscrit dans un domaine interdisciplinaire qui a pour unique objet l'une de ces questions :

Comment peut-on utiliser textes ou paroles en langues humaines en tant que données entièrement appréhendables, analysables et productibles par un programme informatique ?

Ou, plus simplement :

Comment faire « comprendre » une phrase ordinaire à un ordinateur ?

Ce domaine est la *linguistique informatique*[1] ; au carrefour de la linguistique et des mathématiques, de l'informatique et de la philosophie, des sciences cognitives et de l'épistémologie, elle s'est orientée dans deux directions complémentaires.

La première consiste en l'analyse des propriétés des langues humaines, en s'intégrant ou non dans des travaux préalables de linguistique, aux fins de construction de modèles formels rendant compte de l'ensemble des aspects de celles-ci.

La seconde consiste en l'utilisation de résultats partiels facilement applicables dans des programmes informatiques. Cette spécialité, appelée « traitement automatique des langues », est la plus visible de la discipline. Ces dernières années, elle a en effet mis à la disposition des utilisateurs un grand nombre de logiciels de plus en plus aboutis, portant sur la correction orthographique, l'aide à la traduction, la fouille de données…

[1] On utilise fréquemment le terme « linguistique computationnelle », voir [Marcel Cori, 2002].

Le progrès, dans ces deux directions, se révèle cependant long et ardu. En effet, contrairement à de nombreux autres phénomènes (statistiques, modélisation physique…), les processus d'analyse et de synthèse de paroles ou de textes ne sont pas fondés sur des principes arithmétiques déjà connus ; les modèles ont donc, pour la plupart, été créés de toutes pièces. De même, malgré le succès opérationnel des applications ayant été réalisées, celles-ci sont fondées sur quelques propriétés saillantes, faciles à vérifier, mais difficiles à généraliser. Cet ensemble de règles simples et d'heuristiques s'améliore au fil du temps, mais la mise en place d'un programme à couverture totale (tel un traducteur universel) demeure hypothétique, et subordonnée aux avancées de la modélisation.

C'est donc dans ce cadre formel, qui se décline, à son tour, en de nombreuses problématiques, qu'est inscrit ce travail de recherche.

Cette thèse appartient, plus précisément, à la modélisation formelle de la sémantique des langues, c'est-à-dire du « sens » des textes – ou, du moins, d'une représentation utilisable par un programme informatique de ce sens. Dans ce domaine, le terme « sémantique lexicale » regroupe l'étude des mécanismes complexes qui permettent au locuteur de donner une signification à chaque mot employé.

En effet, chaque lexème[2] peut disposer, selon les situations, de différents sens : ce phénomène, la *polysémie*, est une des sources de variabilité des langues.

Comment un programme peut-il choisir, dans une phrase donnée, le sens de chacun des mots la composant ? C'est tout l'objet de la présente thèse : *Modélisation de la sémantique lexicale dans le cadre de la théorie des types*.

[2]Terme dénotant une unité lexicale, hors flexions morphologiques (accords, conjugaisons…), prosodiques ou stylistiques.

Modélisation

Une modélisation d'un phénomène physique, ici la polysémie étudiée par le biais de la sémantique lexicale, consiste en l'élaboration et la validation d'un modèle formel. Ici, il ne s'agit donc pas de savoir comment un locuteur analyse ou produit des phrases, ce qui constitue, par ailleurs, un champ entier de recherche en neurologie et sciences cognitives ; il s'agit de proposer un moyen par lequel un système formel peut choisir, entre les représentations symboliques par lesquelles on peut modéliser les différents sens d'un lexème, celle ou celles qui sont appropriées à un contexte donné. Plutôt que l'exactitude physique, on s'attache donc à l'obtention d'une formulation permettant des simulations satisfaisantes par rapport aux observations.

Si de nombreux linguistes (y compris Noam Chomsky dans [Chomsky, 1955]) considèrent, à juste titre, que la modélisation formelle ne peut permettre seule d'établir une théorie linguistique fidèle au processus réellement employé, elle est en revanche parfaitement adaptée aux buts de la linguistique informatique.

Sémantique

Si l'on s'intéresse ici à la sémantique, c'est afin de permettre à un hypothétique programme capable d'analyser une phrase d'effectuer un traitement arbitraire sur une représentation de son sens. Les applications d'un tel dispositif sont multiples : commande vocale, réponse aux questions, analyse des sous-entendus… Il existe de nombreux systèmes de traitement automatique des langues utilisant ce principe dans des domaines très variés, et qui ne pèchent en général que par l'imprécision de l'analyse de la phrase.

Fondamentalement, les problèmes soulevés par la sémantique sont également plus délicats que ceux d'autres domaines comme la syntaxe. En effet, si des formalismes très puissants ont pu être théorisés pour ce dernier cas, avec une couverture certes imparfaite mais pouvant être étendue à l'ensemble des phénomènes rencontrés, la sémantique butte, elle, sur les manquements du lexique.

Lexique

Lors d'une analyse sémantique est supposé connu un *lexique*, un ensemble de mots associés à une représentation de leur sens, qui permet le calcul de la représentation du sens de la phrase, puis du texte. Or, cette conception naïve ne résiste pas à la polysémie des langues.

Pour la pallier, les systèmes d'analyse sémantique font appel à un procédé inspiré des dictionnaires : pour chaque mot, l'ensemble des sens possibles est détaillé, avec des données telles que des exemples de locutions dans lequel le mot est employé avec chacun de ses sens. Une analyse du contexte, le plus souvent statistique, permet de choisir l'entrée la plus appropriée du « dictionnaire ».

Cette approche, en pratique très souvent utilisée, pose deux problèmes majeurs.

D'une part, elle est empirique, fondée sur des heuristiques dont la couverture n'est assurée que par un travail constant de recherche et de mise à jour des données linguistiques.

D'autre part, l'existence même d'une liste finie de sens associée à chaque mot est sujette à caution ; les théories retenues dans nos travaux supposent le contraire, en se basant, notamment, sur l'étude de mots comportant plusieurs sens fortement liés entre eux.

Dans ce cadre, le lexique devient bien plus complexe qu'une liste associative ou un dictionnaire à entrées multiples : il s'agit d'un ensemble de données et de mécanismes permettant, dans un contexte donné, d'engendrer l'ensemble des sens possibles pour un lexème, et qui sera l'objet principal de ce manuscrit.

Théorie des Types

Le cadre de la *théorie des types* est représentatif de la volonté d'inscrire ces travaux dans un formalisme mathématique et informatique bien défini. Parmi ces formalismes hérités de Russel, ce manuscrit utilise le *Système F* de Jean-Yves Girard pour ses propriétés claires, correspondant au phénomène étudié, et facilement transposable sous forme de programmes. Ce choix préalable, s'il impose des restrictions qui peuvent sembler artificielles, permet de s'assurer d'obtenir un formalisme exploitable, tout en laissant ouvert l'ensemble des possibilités d'interprétation dans différents modèles.

Langue et information

Le but de l'analyse de la langue par des moyens informatiques reste la récolte et le traitement de l'information contenue dans les textes et discours analysés. En complément aux sujets principaux d'études et de recherche, ce manuscrit propose également quelques pistes en ce sens, basées sur une notion de représentations minimales de connaissances et un traitement simple et rapide.

Contenu du présent manuscrit

Après la présente introduction (Ch. 1, p. 13, nous avons organisé le manuscrit en trois parties : une première résumant l'état des connaissances actuelles du domaine, une seconde détaillant notre proposition et une dernière sur les prospectives possibles.

La première partie comporte les chapitres suivants : Sémantique et informatique, Ch. 2, p. 25, qui présente la sémantique formelle, la sémantique lexicale et les très nombreux travaux nous ayant précédé. Modèles de la composition, Ch. 3, p. 49, qui résume les théories de la composition sémantiques et les principes de l'organisation des types. Le comportement logique des phénomènes-cible, Ch. 4, p. 53, qui examine de nombreux exemples qu'il nous faut traiter et leur comportement.

La deuxième partie développe les points suivants : Mécanismes d'ordre supérieur, Ch. 5, p. 73, qui présente les principes d'utilisation du système logique que nous allons employer. Un processus intégré d'analyse des sens, Ch. 6, p. 85, qui détaille notre proposition : système de calcul et d'analyse sémantique. Complétude et apports de l'approche, Ch. 7, p. 93, qui revient sur notre proposition en examinant sa pertinence vis-à-vis des phénomènes à traiter et des autres approches que la notre. Implémentation, Ch. 8, p. 117, décrit les travaux d'implémentation réalisés.

La troisième partie se compose du chapitre Discussions et propositions, Ch. 9, p. 131, qui examine certains des points laissés de côté par notre proposition et des méthodes possibles pour les résoudre.

La conclusion se situe en Ch. 10, p. 139, et les mots-clés et références, p. 143.

Première partie

État de l'art

Chapitre 2

Sémantique et informatique

2.1 Principes de la sémantique compositionnelle

Si la langue est un objet d'études et de recherche depuis l'antiquité, les formalismes récents sont les héritiers de la tradition de Saussure ([Bally and Sechehaye, 1916]). Si les formalisations de Chomsky, Quine ou leurs contemporains se sont attachés au *processus* par lequel le locuteur utilise la langue au quotidien, ce fut Montague qui proposa ce à quoi d'autres avant lui s'étaient refusés : [Montague, 1974] introduit un mode de représentation abstraite et de calcul du sens des phrases. Ce procédé, depuis formalisé et bien connu sous le nom d'*analyse sémantique Montagovienne*, peut être adapté à de multiples usages ; voici son fonctionnement habituel.

2.1.1 L'analyse syntaxique

En utilisant un *formalisme grammatical* (qui peut être une simple grammaire algébrique ou un formalisme légèrement contextuel, voir [Mery et al., 2006]) décrivant les liens syntaxiques entre les constituants de la phrase, on obtient la structure syntaxique de la phrase sous la forme d'un arbre de dérivation conforme, par exemple, à la théorie X-barre (de [Chomsky, 1970]). C'est cette structure syntaxique qui sert de base aux mécanismes compositionnels de la sémantique, en partant du principe que la syntaxe de la phrase sert de guide pour la reconstruction des sens.

2.1.2 La composition

La structure arborescente de dépendances obtenue par l'analyse syntaxique est ensuite convertie en une suite de termes d'un calcul adapté, tel le λ-calcul simplement typé. Les prédicats, qui forment les nœuds de la structure, s'appliquent à leurs arguments suivant l'ordonnancement ainsi calculé.

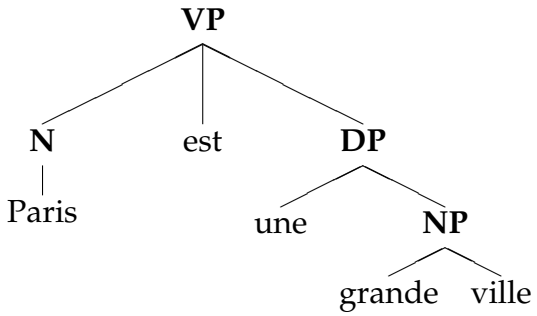
L'analyse sémantique proprement dite consiste ensuite au remplacement des lexèmes par les termes qui leur correspondent dans le lexique (constantes pour les feuilles de la structure syntaxiques, fonctions pour les prédicats), et à effectuer les applications de proche en proche suivant le calcul utilisé (ici, la β-réduction), jusqu'à obtenir une formule logique représentant le sens calculé.

2.1.3 Exemple

Soit la phrase :

Paris est une grande ville

Une analyse classique de la syntaxe de cette phrase est la suivante :



L'analyse sémantique donne la suite de termes suivante :

$$\lambda x^e y^e.(\text{est}^{e\to e\to t})\ \text{Paris}^e\ (\lambda x^e.(\text{une}^{e\to e})\ (\lambda x^e.(\text{grande}^{e\to e})\ \text{ville}^e))$$

Ce qui se réduit en :

$$(\textit{est Paris}\ (\textit{une}\ (\textit{grande ville})))$$

Buts

L'objet de la réalisation de telles formules logiques est leur *interprétation* dans un modèle donné ; la formule elle-même est une représentation du sens de la phrase qui peut être utilisée par un ordinateur, en conjonction avec une base de connaissances, par exemple. Les applications varient selon ce que l'on cherche.

2.2 Problèmes posés par la sémantique lexicale

La *sémantique lexicale* s'intéresse, dans le cadre du processus d'analyse de la phrase décrit précédemment, au sens de chaque mot tel que perçu dans un contexte donné. En supposant connu un mécanisme d'analyse syntaxique permettant de déduire une représentation fonctionnelle de la phrase, et un mécanisme d'analyse sémantique permettant de calculer, suivant cette représentation et une *forme logique* associée à chaque mot, une modélisation du sens de la phrase, la question à laquelle se propose de répondre la sémantique lexicale est la suivante :

Comment, à chaque mot du langage employé dans un contexte donné, associer une forme logique représentant le sens approprié pour ce mot ?

Le mécanisme d'association d'une forme logique telle qu'utilisable dans le formalisme choisi (comme Paris^e) est appelé *lexique*. Et si, pour effectuer une analyse sur des phrases simples, un lexique naïf, obtenu par l'association d'une unique forme logique pour chaque mot (telle que **Paris** := Paris^e), peut suffire, il échoue face à toute forme d'*ambiguïté lexicale*.

Le cas des mots pouvant revêtir, dans différents contextes, de multiples sens, est un problème majeur qui ne dispose pas de solution communément admise, et les avis divergent sur le traitement de ce phénomène, la *polysémie*.

2.2.1 Polysèmes

Dans les langues humaines, la polysémie est banale au point d'être plus souvent la règle que l'exception. N'importe quel dictionnaire illustre cette multiplicité des sens en proposant, très souvent, plusieurs entrées pour un mot donné. Pour appréhender correctement le sens d'une phrase, disposer d'un mécanisme qui choisisse correctement le sens de chaque mot est indispensable…

Ce phénomène est lui-même la somme de nombreux mécanismes. On distingue notamment les cas d'*homophonie* (« après »/ « apprêt ») et d'*homonymie* (« bar » en tant que ressource halieutique / « bar » en tant que débit de boissons), dans lesquelles les différents sens peuvent être distingués facilement par un champ lexical ou même un type syntaxique différent, et celui de la *polysémie logique* (« l'*école* est au bout de la rue »/ « l'*école* est en grève »/ « l'*école* est en vacances »/ ...), dans laquelle les différents sens sont liés par un ensemble de relations logiques.

L'opposition entre homonymie et polysémie est facilement définissable : alors que la première est due à la proximité sonore (*homophonie*) ou graphique *homographie*, la seconde concerne la sémantique individuelle du lexème. L'homonymie peut être levée par de multiples indices : co-texte, catégorie syntaxique... Tandis que la polysémie doit être traitée par une analyse sémantique. C'est cette polysémie logique, appelée également *ambiguïté relationnelle*, qui pose le problème le plus intéressant. En effet, il s'agit moins ici de séparer deux sens bien distincts pour un même mot que d'examiner les mécanismes permettant au locuteur de choisir, dans un contexte donné, l'*aspect* le plus approprié sans en occulter les autres.

2.2.2 Cas concrets

Les phénomènes suivants sont bien identifiés dans la littérature.

Qualia

Une ambiguïté repérée assez tôt est appelée *exploitation des qualia*, en référence aux *qualia* ou $\alpha\iota\tau\iota\alpha$, les « quatre causes »de la philosophie Aristotélicienne. Aristote définissait, dans un texte extrêmement connu et repris depuis lors ([Aristote, 350]), quatre « causes » à chaque action, distinguant l'action elle-même, son déroulement, son auteur et son but. De même, [Moravcsik, 1982] et [Pustejovsky, 1995] ont associé, à un terme donné, quatre *qualia* : le *quale formel*, c'est-à-dire le terme lui-même et l'ensemble de ses propriétés, le *quale constitutif*, c'est-à-dire ce dont le terme est constitué ou ce qu'il constitue, le *quale agentif*, c'est-à-dire ce qui a permis la création du terme, et le *quale télique*, c'est-à-dire ce que ce terme a pour but. Toutes ces relations sont *ontologiques* : elles participent de la connaissance du monde, et non de celle de la phrase (tout en faisant partie de l'espace linguistique : un locuteur compétent d'un langage les utilise quotidiennement).

Ainsi, le terme *épée* peut revêtir les sens suivants :

- *Une épée courte* : l'épée en tant que telle (formel).
- *Une épée bien trempée* : la lame (constitutif).
- *Une épée de maître* : le forgeron (agentif).
- *Une épée efficace* : pour le combat (télique).

Processus et résultat

Une seconde ambiguïté repérée rapidement (par exemple par Quine dans [Quine, 1960]), et détaillée par [Jacquey, 2001] pour quelques cas en français, est celle entre un même terme réalisatif pouvant occuper les valeurs de *processus* ou de *résultat*. Ces termes, dits *déverbaux*, ont fait l'objet de nombreux traitements ; [Grimshaw, 1990], notamment, en donne une analyse détaillée. Cette ambiguïté apparaît dans un exemple longuement repris par [Pustejovsky, 1995], le mot *construction* :

- *La construction dura cinq mois* : processus.
- *La construction est au coin de la rue* : résultat.

[Bassac, 2006] fait état de différentes approches morphologiques permettant de lever cette ambiguïté, avant de se concentrer sur l'apport du lexique génératif à ce problème particulier. Ainsi suggère-t-il des opérations lexicales permettant d'engendrer morphologiquement les déverbaux en donnant à leur *quale* formel le résultat, et à l'agentif, le processus.

Facettes

[Pustejovsky, 1995] et [Asher, 2006] distinguent des ambiguïtés de *facettes* (pour des termes appelés *objets complexes* ou *pointés*) qui font référence, pour un même terme, à des aspects ontologiquement différents.

Le prototype de cette ambiguïté de facettes est le mot *livre*, qui recouvre deux aspects : celui de l'ouvrage, de l'information écrite, et celui de l'objet physique, du volume. Les deux aspects sont très différenciés : l'un est abstrait et obéit aux règles régissant les concepts, l'autre est concret et obéit aux règles physiques. Et on a donc :

- *Un livre lourd* : le livre en tant qu'objet (physique).
- *Un livre intéressant* : le livre en tant que concept (information).

Le principe de [Asher, 2006] est de présenter une opération spécifique, le produit-•, permettant de tenir compte des deux types différents. [Pinkal and Kohlhase, 2000], entre autres, suggèrent *a contrario* qu'une opération d'unification de ces types est suffisante pour ce problème.

2.2.3 Inadéquation de l'approche naïve

La réponse classique à la polysémie est simplement de considérer non pas une entrée lexicale par mot, mais une par sens ; les *lexiques à énumération de sens* (tels les dictionnaires) sont légion.

Les lexiques énumératifs

En effet, en supposant fini le nombre de sens de chaque mot, l'approche d'un dictionnaire ou lexique à énumération de sens est simple : caractériser le comportement de chacun des sens engendré, donner une liste (avec des méthodes analogiques ou statistiques) des contextes dans lesquels chacun apparaît, et associer à chacun une entrée lexicale distincte. Les informations contextuelles de chaque entrée sont utilisées pour sélectionner, parmi les multiples sens d'un même lexème, celle qui correspond le mieux à une utilisation donnée.

Ambiguïté contrastive

Dans les cas d'*ambiguïté contrastive*, en particulier de l'homophonie accidentelle entre deux lexèmes distincts, cette approche est correcte.

Lexème	Type	Description
Bar	*Poisson*	ressource halieutique
Bar	*Etablissement*	débit de boissons

Dans ce cas, où les types sont radicalement différents, le typage du prédicat suffit à donner un résultat non amibigü. Notons cependant que pour prendre en compte un usage supplémentaire du même lexème, dans ce paradigme, il faudrait obligatoirement modifier le lexique, en ajoutant par exemple :

Lexème	Type	Description
Bar	*Unite*	pression atmosphérique

Ambiguïté relationnelle

Lorsque l'ambiguïté porte sur deux sens en rapport logique, l'approche énumérative fonctionne jusqu'à un certain point : elle reste limitée à l'énumération donnée dans le lexique, qui peut être longue et fastidieuse :

Lexème	Type	Description
Rapide	$Animal \to t$	un animal se déplaçant rapidement
Rapide	$Vehicule \to t$	un véhicule capable d'une vitesse jugée rapide
Rapide	$Evenement \to t$	un événement se déroulant dans un court laps de temps

Une telle énumération est difficile à rendre exhaustive, et ne peut pas tenir compte des évolutions et nouveaux usages de la langue : comment rendre compte du sens de *rapide* dans *un téléphone rapide* pour un lexique précédent la téléphonie mobile de troisième génération ? C'est sur la base de ce constat que nous pouvons dire avec [Pustejovsky, 1995] que les lexiques énumératifs ne sont pas idéaux pour tenir compte des ambiguïtés relationnelles, et c'est tout l'intérêt d'un lexique génératif basé sur des inférences logiques entre éléments lexicaux.

2.3 Théories préliminaires

2.3.1 Connaissances préalables

L'idée que l'analyse de la langue soit associée à des connaissances dites « d'arrière-plan » peut être attribuée à [Searle, 1979]. Searle y exprime la nécessité de connaissances préalables au langage pour la compétence langagière, qui en seraient indissociables.

2.3.2 Les rôles sémantiques

Les grammaires de cas

Fondée sur les notions classiques de *cas*, la *grammaire de cas* est une famille de formalismes développés pour rendre compte de l'inflection casuelle (notamment en Grec ou Latin) ou de la position des mots faisant correspondre les mêmes fonctions grammaticales dans les langues non-casuelles, ainsi que de leur impact sur la syntaxe des phrases.

Se basant sur cette longue tradition, et en réaction à la vision transformationelle de la grammaire proposée par Chomsky, [Fillmore, 1965] développe une version des grammaires de cas dans laquelle les relations sémantiques auront une priorité. Cette dernière notion, encore mal définie, s'appuie sur les *rôles sémantiques* de l'Objet, de l'Agent et du Datif, qui seront modifiés à des nombreuses reprises dans les travaux subséquents de Fillmore (voir [Anderson, 2006] pour une remise en perspective de l'ensemble de ces travaux).

Cas et structure profonde

Les relations explorées dans la grammaire de cas de Fillmore sont un guide pour la structure de la syntaxe qui doit rendre sans objet la « structure profonde » des grammaires transformationelles-génératives de Chomsky, et remplacer cette dernière. De fait, la poursuite de ces travaux conduiront les tenants de la grammaire générative à intégrer une certaine vision des rôles sémantiques dans la structure de leurs grammaires. La plus grande contribution de Fillmore, dans ce cadre, est d'avoir initié de nombreux travaux et une approche incontournable à l'influence des relations sémantiques sur la phrase, qui sera prise en compte dans la majorité des théories linguistiques subséquentes.

Les rôles thématiques

Les grammaires de cas de Fillmore, ainsi que les relations lexicales étudiées dans [Gruber, 1965], seront généralisées pour obtenir les θ-rôles, ou *rôles thématiques*. Les définitions de Fillmore décrivent les rôles suivants :

- *L'agent* est le cas de l'élément typiquement animé, perçu comme instigateur de l'événement identifié par le verbe.
- *L'instrumental* est le cas de la force ou de l'objet non-animé qui est impliqué causalement dans l'événement identifié par le verbe.
- Le *datif* est le cas de l'élément animé qui est affecté par l'événement identifié par le verbe.
- Le *factitif* est le cas de l'objet ou de l'être animé qui résultent de l'événement identifié par le verbe.
- Le *locatif* est le cas du lieu ou de l'orientation spatiale de l'événement identifié par le verbe.

- L'*objectif* est le cas le plus neutre du point de vue sémantique : c'est le cas porté par un nom identifiant un élément affecté par l'événement décrit par le verbe.

Notons que ces notions, encore définies par des cas, sont centrées sur l'événement référencé par le prédicat de la proposition analysée. De plus, Fillmore indique ici que « l'ajout de plusieurs cas supplémentaires sera probablement nécessaire » : cette liste n'est pas exhaustive, et a été maintes fois modifiée par la suite.

Quoi qu'il en soit, ces θ-rôles ont fait l'objet d'une intégration au sein des grammaires transformationnelles-génératives. Ils ont été d'un impact important : malgré l'absence de définitions formelles, l'instabilité des valeurs accordées à chaque rôle et même de leur nombre, et les évolutions constantes apportées par de multiples travaux, ils sont désormais considérés comme une composante intrinsèque de toutes les grammaires génératives actuelles.

Cependant, cette grande variabilité (au fil du temps, Cook suggérera une hiérarchie de cinq cas, tandis que, chez Fillmore, le *datif* sera partiellement transformé en « expérienceur »…) et surtout le manque de formalisme parasitant les définitions amènera Dowty, entre autres, à se préoccuper de la validité épistémologique de ces θ-rôles. Dans [Dowty, 1989], il s'interroge sur la pertinence de ces notions, et en vient à redéfinir les rôles comme autant de prédicats sur la sémantique de la phrase.

Une base pour l'analyse sémantique

En général, les cas ou rôles thématiques sont l'expression d'une couche supplémentaire sur une vision simplement fonctionnelle de la syntaxe : aux notions de sujet, verbe, objet, on substitue des relations d'origine primairement sémantique d'agent ou d'objectif, autour d'un prédicat. Cette vision permet de s'affranchir de certaines particularités idiosyncratiques de la langue et de proposer une théorie de la prédication basée sur la sémantique.

Il s'agit d'une étape vers le développement des *qualia* de Pustejosky et des diverses structures présentées dans [Pustejovsky, 1995], car les rôles thématiques décrivent différentes fonctions sémantiques associées aux éléments de la phrase, ainsi que les opérations associées qui affectent aussi bien la syntaxe que la sémantique du langage. De même, ces rôles sont la fondation de l'approche de [Gupta and Aha, 2003], qui associe une structure orientée objet complète à un ensemble de propriétés qui sont des rôles thématiques choisis et redéfinis.

2.3.3 Réification des événements

D'autre part, [Davidson, 1967] a développé une branche de l'analyse sémantique (souvent appelée *sémantique davidsonnienne*) caractérisée par la *réification des événements*, c'est-à-dire par la représentation explicite des événements en tant que faits de base pour les verbes d'action. [Parsons, 1989] fait également état de ce principe, distinguant entre différentes sortes d'événements qui ont des répercussions sur la grammaticalité même de la phrase. Ces travaux serviront de base à la structure événementielle de [Pustejovsky, 1995] et à la logique temporelle.

2.3.4 Le Lexique Génératif

Le lexique génératif, décrit en premier lieu dans [Pustejovsky, 1991] et détaillé par la suite dans [Pustejovsky, 1995], constitue la fondation de nos approches. Son but : représenter les informations minimales à partir desquelles, dans un contexte donné, un même mot peut prendre différents sens liés. Il détaille ainsi de nombreux cas connus d'ambiguïté relationnelle, ainsi que les usages créatifs des mots.

Principe de co-compositionnalité

Le principe directeur du lexique génératif est une surcouche à l'analyse grammaticale de Montague, qui vient se greffer sur un arbre syntaxique d'une phrase déjà construite. Le lexique génératif distingue alors pour les différents lexèmes des *types raffinés*, établissant une distinction ontologique entre les différents sens d'un même mot. Ainsi, le prédicat *engagé* attendra un argument de type *personne*, et un *article engagé* présente alors une contradiction de types, résolue par les mécanismes de *coercition de type* ou d'*accomodation*.

Dans la *coercition*, un conflit de types est résolu par la présence, dans la structure lexicale de l'argument, du type souhaité. On remplace alors l'argument par la composante à laquelle est affectée ce type.

Dans l'*accomodation*, le même conflit est résolu par la modification des types concernés, via une certaine relation de sous-typage qui sera explicitée plus loin. Les deux opérations respectent un principe de *co-compositionnalité* : les informations du prédicat et de l'argument contribuent tous deux à la réalisation du résultat de leur composition.

Ce processus est subtil et permet la participation à l'analyse sémantique d'un lexique très riche, aux multiples dimensions. Il est précisé, au travers de quelques exemples, la manière dont une analyse logique ferait appel aux mécanismes de types ; malheureusement, il est difficile de systématiser ces exemples, et de multiples aspects parfois contradictoires sont pris en compte.

Structures lexicales enrichies

Pour permettre aux mécanismes d'actions sur les types raffinés mis en place d'être effectifs, Pustejovsky enrichit la structure des entrées lexicales. L'organisation des données devient orientée vers l'application de ces mécanismes.

Ainsi, chaque entrée regroupe :

- Une structure argumentale. Il s'agit du λ-terme du lexème, donné avec les types raffinés de ses arguments, ainsi qu'une indication pour les *arguments optionnels* et les *arguments par défaut*.
- Une structure événementielle[1]. Elle décrit, s'il y a lieu, l'ensemble des événements et sous-événements et leurs relations qui concernent le terme, et est notamment utilisée pour l'ambiguïté processus/résultat.
- Une structure de *qualia*. Elle décrit, pour chacun des quatre *qualia*, les termes en relation et leurs types raffinés, s'il y a lieu.
- Une structure d'héritage. Elle permet la factorisation du lexique *via* l'hyponymie des termes.

Voici un exemple d'entrée lexicale pour *épée*, dans la présentation usuelle de Pustejovsky, similaire à une structure de traits :

$$\left[\begin{array}{rcl} \textbf{epee} & & \textit{arme} \\ \mathsf{ARG} & = & \lambda x.(\text{epee } x) \\ \mathsf{EVENT} & = & \\ \mathsf{QUALIA} & = & \left[\begin{array}{rcl} \mathsf{FORMAL} & = & \textit{epee} \\ \mathsf{CONST} & = & \exists m^{\varphi},\ ((\text{metal } m) \wedge ((\text{partie } x)\ m)) \\ \mathsf{AGENT} & = & \exists y^{A},\ ((\text{forge } x)\ y) \\ \mathsf{TELIC} & = & \exists z^{Evt},\ (\text{combat } z)\ x) \end{array}\right] \end{array}\right]$$

[1] Bien que la logique temporelle donnée par Pustejovsky permette une grande variété de comportements et soit assez complète, nous ne nous intéresserons pas à cet aspect précis de la sémantique lexicale, préférant les cas de polysémie.

Chaque aspect de la structure dispose d'éléments informatifs sur son type et les éventuels termes à associer dans une opération logique. Les variables introduites dans chaque champ peuvent ainsi être unifiées avec ceux d'autres champs d'autres entrées, si les types sont compatibles.

Facettes multiples

De plus, le lexique génératif introduit la notion *d'objets complexes*, qui disposent de deux types ou plus. Ces lexèmes, également appelés •-objects (*dot objects*) reflètent des mots qui disposent de plusieurs sens fortement liés mais apparemment incompatibles : des mots à *facettes multiples*. On note un mot aux facettes A et B comme étant de type $A \bullet B$; ainsi *livre* est considéré comme *simultanément* un objet physique et le contenu informationnel associé.

Récapitulatif des modes opératoires

Les opérations génératives, permettant l'utilisation des mécanismes spécifiques au lexique génératif, sont les suivantes, toujours selon [Pustejovsky, 1995] :

Coercition
Lorsqu'un prédicat est sensé porter sur un argument d'un certain type, et qu'il est appliqué à un terme d'un type différent, s'opère l'opération de *coercition* : en faisant appel à des informations lexicales, le type de l'argument est changé, modifié, transformé, forcé. Selon les informations utilisées, on distingue :

- *Exploitation de Qualia* : l'un des *qualia* de l'argument est utilisé en lieu et place de ce dernier. Exemple : *Cet article est irrévérencieux.* Le prédicat fait appel à une personne, et on dispose d'un écrit en argument ; on utilise l'*agentif*, qui est de type personne, et on pourrait gloser le résultat ainsi : *L'auteur de cet article est irrévérencieux.*

- *Exploitation de facette* : une des facettes de l'argument est utilisée au lieu de l'ensemble des facettes. Exemple : *Ce livre est lourd.* Selon les termes du lexique génératif, l'argument dispose de deux facettes, objet physique et contenu informationnel, alors que le prédicat ne fait appel qu'aux propriétés physiques. Le résultat de l'opération peut être glosé ainsi : *Ce livre, en tant qu'objet physique, est lourd.*

- *Accomodation* : il est fait appel, plutôt qu'à l'argument tel quel, à une propriété générique de ce dernier. Exemple : *Cette Honda est rouge.* L'argument est très spécifique et dispose d'un type bien précis dans la hiérarchie (voir plus loin), alors qu'il n'est fait appel qu'à une propriété des objets physiques. On peut gloser le résultat de l'opération ainsi : *Cette Honda, qui est une voiture, c'est-à-dire un véhicule, donc un artefact, donc un objet physique, est rouge.*

Spécification
De multiples prédicats sont définis dans le lexique comme ayant un type générique, sous-spécifié, en argument. Lorsqu'ils sont employés avec un certain argument, ce type vague est spécifié par l'énonciation, et certaines informations lexicales sont alors complétés par l'argument reçu. On distingue, entre autres, les opérations suivantes :

- *Co-composition* : un verbe au sens vague est utilisé avec un argument précis, sa structure de *qualia* et sa structure événementielle sont alors complétées. Exemple : *I hammered the sword.* (activité), *I hammered the metal flat.* (changement d'état).

- *Liage sélectif* : un adjectif ou un adverbe pouvant concerner de multiples aspects voit le *qualia* adapté sélectionné lors de la prédication. Exemples : *Une voiture rapide*, *Un bon couteau* : télique.

2.4 L'Ontologie et le lexique

Les types raffinés de [Pustejovsky, 1995] sont un cas particulier de TY_n, une logique de types à n sortes. Le principe directeur de cette approche est de tendre vers d'une *ontologie de types* fondée sur les informations inhérentes à chaque terme, plutôt que d'utiliser l'approche de Montague qui est une classification des termes en fonction de leur applicabilité aux autres termes de la phrase.

2.4.1 Pourquoi une ontologie ?

Une ontologie, d'après [Chandrasekaran et al., 1999] (entre autres) est une représentation du monde tel que nous le percevons, et de la vérité des rapports entre toutes choses qui s'y trouvent. Si nous nous y intéressons et si [Pustejovsky, 1995] suppose un tel élément, ce n'est pas pour disposer d'un objet philosophique universel, mais bien pour s'appuyer sur un

outil de raisonnement sur le langage. [Smith, 2003] rappelle, à juste titre, que la construction d'une ontologie universelle et exhaustive est une utopie ; le propos est ici de considérer une ontologie restreinte aux termes du langage et aux informations qu'ils véhiculent. Cette ontologie est alors un complément au lexique, et non pas un objet à part.

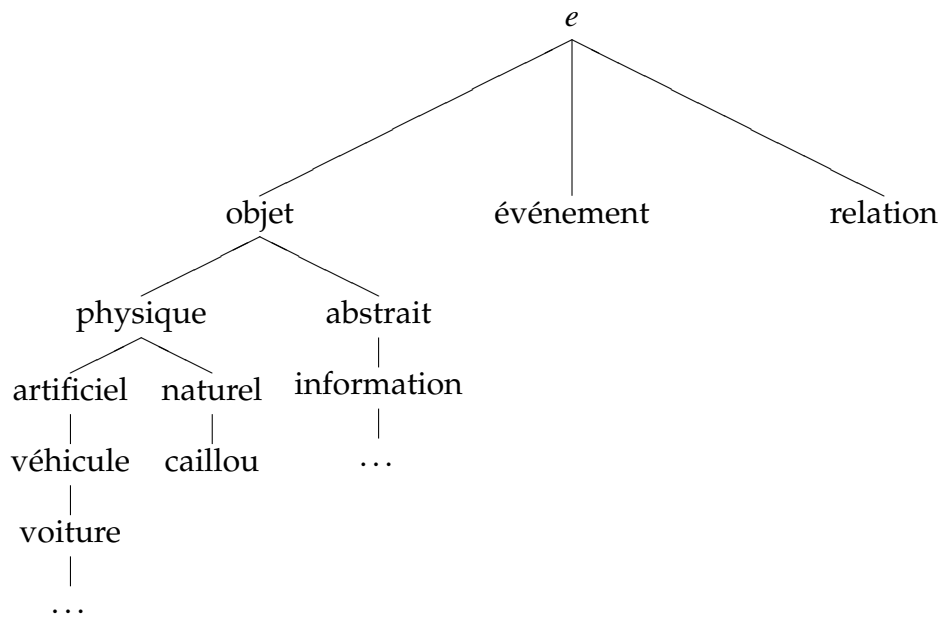
2.4.2 Les types ontologiques du lexique génératif

Dans [Pustejovsky, 1995], comme le rappellent [Nirenburg et al., 1995], Pustejovsky emploie implicitement une *ontologie de types*, formée par la *structure d'héritage* du lexique génératif.

Concrètement, en lieu et place de la logique montagovienne à une sorte, qui considère le type des entités (e) et celui des valeurs de vérité (t), on dispose d'une grande variété de types venant « remplacer » celui des entités.

Ces types sont organisés en arborescence, du plus général au plus particulier, et décrivent précisemment les différences de comportement que le lexique génératif se propose de capturer.

On pourrait ainsi avoir quelques types comme :



Le but d'une telle ontonlogie (simplement, encore une fois, supposée par Pustejovsky et non détaillée) est de couvrir l'ensemble du lexique ; chacun des nœuds et feuilles de cet arbre sont des types.

2.4.3 Conséquences pratiques

L'héritage entre les structures lexicales, présenté sous cette arborescence, donne immédiatement accès aux relations d'*hyponymie* et *hyperonymie*.

Ainsi, la hiérarchie précédente nous donne directement le fait que *voiture* soit un hyponime de *objet physique*, et on peut donc paraphraser les prédications comme *une voiture rouge* comme étant *un objet physique (qui est une voiture) de couleur rouge*. De même, *objet physique* est un hyperonime de *caillou*, on pourrait donc inférer *par exemple un caillou* à partir de *un objet lourd*.

La relation d'hyponymie notamment donne accès à de nombreuses informations qui peuvent être factorisées au sein de cette hiérarchie.

2.5 Critiques et compléments du lexique génératif

Critiques de courant

Si la proposition générale contenue dans [Pustejovsky, 1991] et les travaux subséquents mérite amélioration, elle a également été fortement critiquée sur le fond. Ainsi, [Fodor and Lepore, 1998] considèrent que la proposition est vide de sens, en ce que le lexique doit être atomique et le sens dénotationnel. Ce point de vue a fait l'objet de réponses acerbes de l'auteur, notamment [Pustejovsky, 1998]. De manière plus mesurée, [Blutner, 2002] argue de l'impossibilité d'opérer une distinction contextuelle lors de l'analyse sémantique, renvoyant à une analyse pragmatique postérieure les mécanismes subtils du lexique génératif.

Controverse : la nature du lexique

La critique formulée par [Fodor and Lepore, 1998] est d'importance centrale pour notre propos, il nous donc faut l'examiner en détail, ainsi que les contre-arguments qui lui sont opposés. En effet, admettre la nature atomique et purement référencielle du lexique proposée par Fodor et Lepore reviendrait à renoncer à toute analyse du sens du langage basée sur la sémantique lexicale, et rendrait donc l'ensemble des présents travaux caducs ; nous devons donc défendre la pertinence de notre proposition par rapport à cette thèse.

La critique en elle-même est spécifiquement orientée contre [Pustejovsky, 1995]. Deux publications y répondent : [Pustejovsky, 1998], qui reprend les arguments visant des constructions spécifiques du lexique génératif et donne des explications basées sur de multiples exemples, et [Wilks, 2001], qui s'oppose aux critiques de fond et constitue une réfutation forte de chaque argument présenté.

Sens et inférences
Le premier sujet à controverse est le refus par Fodor et Lepore de la présence d'inférences dans le sens des mots (la relation *un chien est un animal* est une partie du sens du mot *chien*). L'association d'inférences informationnelles au sens d'un mot est le fondement du lexique génératif, mais aussi de tout un pan de théories d'analyses du langage naturel et de l'intelligence artificielle ; c'est pour nous une donnée essentielle.

Sémantique à rôle informatif
[Fodor and Lepore, 1998] définissent comme un courant de la linguistique, de la philosophie et des sciences cognitives la sémantique à rôle informatif (*Informative Role Semantics*) le postulat que le sens d'une expression linguistique contient une partie de ces relations inférentielles. Les auteurs rejettent de façon absolue ce courant (tout en reconnaissant que ce postulat est presque universellement admis), sans donner de véritable justification autre qu'un *a priori* négatif, et indiquent qu'ils considèrent comme caduque toute théorie en découlant, dont le lexique génératif (et, par conséquent, le présent manuscrit). Comme [Pustejovsky, 1998], nous considérons que ce point de vue est très pessimiste en ce qu'il considère qu'il ne sera jamais possible de donner une explication satisfaisante à la créativité du langage ; comme [Wilks, 2001], nous pensons que le sens d'un mot contient bien des relations inférentielles de ce mot par rapport aux autres, sans quoi le lexique n'offrirait aucune explication, aucun moyen de comprendre un mot ou d'appréhender son sens autre que par ostension, c'est-à-dire le mot lui-même.

Pour le moins, nous soutenons que l'hypothèse d'une sémantique à rôle informatif ne peut être rejetée sans aucune justification.

Notion de nécessité

[Fodor and Lepore, 1998] développent ensuite une notion très mal définie sur laquelle ils s'appuient pour rejeter les exemples de sens proposés par Pustejovsky : celle de la « nécessité » de la présence d'une inférence dans le sens d'un mot. Ils souhaitent établir que le lexique doit être exempt des inférences qu'ils considèrent comme « non essentielles » pour la compréhension d'un mot et en donnent de très nombreux exemples, ce qui est d'autant plus facile pour eux parce que cette « nécessité » n'est formellement définie nulle part, mais surtout parce qu'ils considèrent que le sens d'un mot est constitué de sa dénotation, sans préciser cette dernière, ni un modèle d'interprétation pour cette sémantique.

Ainsi, les auteurs nous signifient que l'expression *vouloir une cigarette* ne pourrait pas disposer d'une glose par défaut qui serait *vouloir fumer une cigarette*, car il est possible, dans certains contexte, que cette expression signifie *vouloir posséder une cigarette* ; la relation de télique pour *cigarette* (qui est l'association avec le verbe *fumer*) devrait donc ne pas figurer dans le lexique.

Cette approche, justifiée par une longue discussion et un algorithme dont les problèmes sont fort bien mis en lumière par [Wilks, 2001], participe d'une mauvaise foi évidente, en faisant dire à [Pustejovsky, 1995] ce qu'il ne dit pas (ce dernier admet parfaitement que la glose ici donnée puisse être, ou non, permise par le contexte ; c'est toute l'intérêt de cette thèse). Elle est de plus particulièrement fallacieuse, car elle considère que le sens d'un mot ne peut s'accomoder d'une relation qui disposerait d'un contexte particulier la rendant invalide : il nous semble naturel de défendre l'assertion inverse, qui est que toutes les relations qui disposent d'un contexte les rendant valides pour un mot donné participent au sens de ce mot. Ainsi, nous souhaitons soutenir la thèse qu'un locuteur habitué au français puisse savoir que *un chien est un animal*, tout en concevant l'existence de *un chien en peluche*.

Sémanticité (élégance)

Un apport intéressant de [Pustejovsky, 1995] est la notion de « sémanticité », ou élégance sémantique ; elle explique pourquoi certaines expressions comme *manger un livre*, *commencer un dictionnaire*, *une bonne pierre*… paraissent étranges, non naturelles ou discutables pour le locuteur, parce que l'analyse de ces expressions met à jour un conflit de types non résolu. Fodor et Lepore considèrent que cette notion n'a aucune valeur, car chacune de ces expressions dispose de contextes qui la rendent valide ; nous considérons que, là encore, cet argument n'a pas lieu d'être.

En effet, [Pustejovsky, 1995] ne considère en aucun cas qu'il ne puisse exister de contextes rendant valides ces expressions ; au contraire, il indique que ces expressions données telles que *nécessitent* un tel contexte pour être comprises. [Pustejovsky, 1998] donne de nombreux exemples supplémentaires de telles constructions « asémantiques » (nous dirions plutôt « inélégantes », car ces phrases au contenu sémantique difficile à cerner peuvent être utilisées par un auteur talentueux ou osé pour induire un certain effet de style). Et notre propre analyse est que ces expressions font appel à une transformation indéfinie, qui peut être apportée par le contexte.

Quelle limite ?

La présence d'inférences dans le sens de chaque mot pose un autre problème à Fodor et Lepore : celui de savoir combien d'inférences doivent être contenues dans ce sens. L'agument est que l'on puisse très bien concevoir ce qu'est un *cercle* sans savoir que *un cercle n'est pas un carré*, et que cette dernière inférence n'est donc pas présente dans le lexique.

Cet argument est fallacieux en ce qu'il considère que le sens d'un mot est figé, unique ; c'est à la fois ne pas comprendre le principe même du lexique génératif (dans lequel le sens des mots évolue) et, surtout, ignorer complètement tout processus d'acquisition du langage.

Il nous semble raisonnable de penser que le locuteur, au fil de sa vie, puisse apprendre le sens de nouveaux mots, et enrichir le sens des mots qu'il connaît déjà par l'apprentissage. Supposons qu'un enfant ignore la relation entre une *cigarette* et *fumer*, tout en connaissant l'objet *cigarette* ; il pourra, par la suite, apprendre cette relation par le biais de questions telles *à quoi sert ... ?*. Il aura alors enrichi son lexique par cette relation, se rapprochant du lexique avec informations riches de Pustejovsky et s'éloignant du lexique sans aucune information de Fodor et Lepore.

Dénotation et représentation

Pour [Fodor and Lepore, 1998], le sens d'un mot est atomique et réduit à sa dénotation. Outre le fait, déjà mentionné, qu'il n'est jamais fait référence à un quelconque modèle pour cette dénotation (et qu'alors, comme le souligne longuement [Wilks, 2001], la correspondance sens-dénotation est vide de tout apport informatif), cette hypothèse donne lieu à des arguments totalement fallacieux sur une différence supposée entre dénotation et représentation.

Par exemple, lorsque [Pustejovsky, 1995] analyse les événements (c'est-à-dire les mots comme *chanter*, *commencer*, *rencontre*... qui font tous ré-

férence à des événements), il fait appel à diverses propriétés : il existe deux sous-événements, l'un étant mis en exergue par rapport à l'autre … etc. Pour Fodor et Lepore, il est inimaginable que ces affirmations soient vraies : il n'existe pas de propriété faisant d'un sous-événement quelque chose de plus important qu'un autre, et il existe bien plus de sous-événements que deux, suivant la structure physique de l'univers.

Ces arguments se trompent entièrement de domaine.

Si, en effet, on peut dire que la physique peut déterminer un nombre très important de parties à un événement (autant que d'instants de Plank), jamais Pustejovsky ni aucun autre sémanticien n'y fait référence, ni ne le peut. Chacune des affirmations est faite sur le *langage*, et non la nature physique de l'univers : pour parler d'une *cigarette*, il est totalement à exclure que la bonne sémantique soit l'ensemble des particules et relations entre les particules de l'objet concerné. Nous parlons de sens dans la langue, et donc, forcément, de représentations symboliques ; l'ensemble des arguments de [Fodor and Lepore, 1998] fondés sur cette distinction supposée entre représentation et dénotation est absurde et dangereuse. Elle revient à dire qu'un mot *est* ce qu'il représente, et que nous ne pouvons pas parler d'un objet du monde sans en avoir une parfaite connaissance ; c'est évidemment faux (le simple fait que nous puissions écrire le mot *infini* est une preuve), et cela contredit un argument donné dans le même article et vu précedemment, la limite au nombre d'inférences contenues dans le sens d'un même mot.

Sens en composition, sens en isolation

Un autre argument avancé par [Fodor and Lepore, 1998] est que le sens *intrinsèque à un mot* est différent du sens *du même mot composé avec un contexte donné*. Cet argument est faux : il est invérifiable.

En effet, nous pourrions admettre ou rejeter cette hypothèse sans qu'elle n'ait aucune influence sur quelque partie que ce soit du reste de l'analyse proposée par Pustejovsky.

Si nous l'admettons ou non, le sens d'un mot étant, pour Fodor et Lepore, cette mystérieuse dénotation qui n'est pas employée dans la composition, le fait de le modifier ou non n'apportera rien à l'analyse d'un texte, et sera entièrement invisible puisque nous ne sommes pas sensés avoir accès à cette connotation.

Au contraire, le sens d'un mot tel que donné par [Pustejovsky, 1995] n'a qu'une utilité : savoir quel sera le sens d'une expression composite dans laquelle ce mot sera utilisé. L'étudier en isolation est un exercice entièrement vain : le lexique est sensément génératif et compositionnel, il ne dispose d'aucun impact pour un mot isolé.

Utilité du lexique
L'argument principal de [Wilks, 2001] pour réfuter [Fodor and Lepore, 1998] est simple et direct : il s'agit de donner, suivant les deux hypothèses, l'utilité d'un lexique.

Pour le lexique proposé par Pustejovsky, ou n'importe quelle théorie s'inscrivant dans le courant défendu par Wilks, le lexique donne, pour chaque mot, une *explication* : un ensemble de relations avec d'autre mots, au minimum, permettant à un locuteur d'*apprendre* des informations sur le mot concerné.

Pour le lexique atomique de Fodor et Lepore, la sémantique d'un mot est sa signification, ou, pour reprendre l'exemple donné : *le sens de "chien" est CHIEN* (*chien* étant le mot, *CHIEN* sa dénotation). Selon Wilks, cette analyse revient à dire que le lexique associe à chaque mot le même mot ; cela revient à dire que le lexique est vide de sens (le titre de [Fodor and Lepore, 1998], *The emptiness of the lexicon*, semble signifier que les auteurs sont d'accord avec cette analyse). Le lexique serait alors inutile, et on pourrait dire, en appliquant le rasoir d'Ockham, que le lexique n'existe pas, ce qui serait préjudiciable aux éditeurs de dictionnaires.

Pour nous, cet argument suffit à invalider totalement l'hypothèse de Fodor et Lepore : pour ne prendre qu'un exemple, la notion de lexique (un lexique riche en informations) est centrale à la compilation et à l'interprétation de programmes. Qui pourrait soutenir, en informatique, que les programmes n'existent pas ou n'ont aucune utilité ?

Conclusion : étude du langage
Nous considérons pouvoir rejeter totalement les critiques fondamentales de [Fodor and Lepore, 1998] sur le contenu du lexique. Nous avons donné de nombreux arguments, étayés par plusieurs publications, en ce sens. Mais pour nous, l'argument principal est le suivant : notre analyse ne porte pas sur la philosophie ontologique, mais sur une analyse, au moyen de l'informatique, de textes écrits dans les langues humaines. Les efforts en ce sens, tous basés sur des lexiques contenant une information plus ou moins riches, permettent d'ors et déjà d'aboutir à des résultats concrets ; de la même manière, nous donnons dans le présent manuscrit un exemple d'implémentation d'une théorie basée sur la sémantique lexicale.

Les lexiques permettent objectivement de résoudre certains problèmes posés par l'analyse des langues ; dans ce cadre, ils ne sont donc pas inutiles.

Les opérations lexicales

Dans [Copestake and Briscoe, 1991] et quelques travaux supplémentaires, le lexique génératif est amendé pour tenir compte d'opérations supplémentaires qui sont lexicalisées, telles *Grinding* et *Packaging* (deux opérations destructrices consistant à effectuer un changement structurel sur l'objet représenté par le lexème, respectivement en le destructurant ou en l'amalgammant). Ainsi, l'opération consistant à cuisiner une plante ou un animal peut permettre de donner des phrases telles *Ce mouton est délicieux.*, en employant une opération de transformation d'animal en nourriture qui ne corresponde pas entièrement aux structures de [Pustejovsky, 1995] mais qui y est analogue.

Critiques de forme

De nombreux auteurs , ainsi que nous-mêmes, opposons un autre point de vue à [Pustejovsky, 1995] : tel qu'il est présenté, il n'est en effet pas formalisé. Tout au moins, l'ensemble des opérations ne s'intègre pas dans un cadre existant, ni ne propose son sien propre, et il n'est donc pas utilisable en tant que fragment d'analyse ou de génération d'un texte. La théorie reste fondée, mais isolée des préoccupations des autres canevas du traitement du langage. C'est sur ce point que ce sont focalisés de nombreux travaux postérieurs, dont celui-ci.

2.6 Les formalisations récentes

Plusieurs travaux ont cherché à expliciter les opérations employés dans [Pustejovsky, 1991], sans jamais être totalement satisfaisants du point de vue de l'analyse du langage naturel.

Le lexique génératif

[Pustejovsky, 1995] lui-même cherche à poser plusieurs opérations sous forme de déductions logiques : coercition de types, accomodation, etc. Le problème est que les opérations présentées ne sont pas basées sur un système bien fondé. Elles constituent une explication complète et exhaustive des phénomènes étudiées, mais ne sont pas utilisables telles quelles.

Les glissements de sens

[Nunberg, 1993] propose une théorie des glissements de sens. Dans cette théorie, un prédicat peut changer sens pour s'adapter à un argument qui n'est pas du type attendu. L'approche en elle-même est très intéressante et permet de traiter des cas « oubliés » par le lexique génératif, comme *Je suis garé en double-file.*, où le linguiste peut légitimement arguer du fait que *se garer* et non pas *je* contribue l'information permettant d'inférer le véhicule de la personne. Cependant, elle ne propose aucune formalisation logique dans un cadre d'analyse ou de génération du langage.

Feature Logic for Dotted types

[Pinkal and Kohlhase, 2000] est, à notre connaissance, la première proposition sérieuse d'intégrer les opérations du lexique génératif, et tout spécifiquement ses « types pointés », dans un cadre formel : ici, la logique des enregistrements. Malheureusement, outre les questions posées par ce choix lui-même, peu de retours ont été faits sur ce premier article ; [Jacquey, 2001] a étudié en détails cette proposition, et [Cooper, 2007] reprend une telle approche, mais les auteurs originaux du lexique génératif n'ont que peu abondé pour ou contre cette logique qui reste, en l'état, difficilement utilisable (la traçabilité des opérations, surtout, posant problème).

The metaphysics of words in context

Pour répondre aux critiques et poursuivre la formalisation de son propre travail, Pustejovsky s'est associé à Asher pour un projet ambitieux : la réalisation d'une logique complète correspondant aux contraintes du lexique génératif. [Pustejovsky and Asher, 2000] est la première d'une série de publications sur le même thème, visant à la formalisation d'une telle logique ; Asher poursuit encore ces travaux à l'heure actuelle. Le principe est de formaliser chacune des opérations esquissées dans [Pustejovsky, 1995], avec un système de types cohérent.

Le système résultant, cependant, a fait l'objet de nombreuses critiques et n'est pas encore praticable. Des problèmes de notation et des opérations très complexes font de cet effort un ensemble prometteur, mais perfectible.

À l'heure actuelle, les travaux se poursuivent autour d'une approche complète fondée sur une nouvelle logique de types, présentée dans [Asher, 2011] ; cette théorie ayant été développée parallèlement à la présente thèse, nous nous attacherons à la présenter en détails plus loin dans ce manuscrit, et à la discuter.

Travaux ponctuels

S'appuyant sur [Pustejovsky, 1995] et élaborant une architecture complexe dans un paradigme orienté objet, Kalyan Moy Gupta a proposé des versions opératoires et fonctionnelles de la sémantique lexicale. Dans [Gupta and Aha, 2003] puis [Gupta and Aha, 2005], il donne ainsi deux exemples de travaux basés sur des « ontologies de sous-langage », autrement dit des ontologies très spécialisées pour un domaine.

Dans [Marlet, 2007], Renaud Marlet essaie de combler le vide entre la sémantique compositionnelle et le lexique génératif. Cette proposition est solide et donne l'implémentation en sémantique compositionnelle de deux opérations lexicales, la coercition de types et le liage sélectif. Le calcul est complet, mais les travaux restent préliminaires, car ils laissent de côté de multiples opérations lexicales et qu'ils présupposent une implémentation du lexique génératif.

[Saba, 2007] propose, de son côté, une approche différente : en constatant que l'information lexicale est nécessaire à l'analyse d'une phrase et que des structures de données riches fondées ontologiquement, proches du lexique génératif, sont pertinentes pour cette information, Saba signifie qu'un tel canevas ne peut être construit à partir de rien. Il propose un système *d'apprentissage* de telles informations du corpus. Les arguments sont intéressants et nous pouvons comprendre cette approche, mais la proposition est trop restreinte par rapport aux problèmes logiques connus pour être utilisée telle que.

De nécessaires clarifications

Dans l'ensemble de ces travaux manque encore une solution réalisable, donnant une analyse complète à partir de la syntaxe de la phrase. Pour permettre de la réaliser, nous nous proposons non pas d'intégrer dans un quelconque système fragmentaire l'ensemble des propositions de Pustejovsky et de ses suivants, mais de partir d'une analyse bien connue – le calcul sémantique de Montague, qui, dans [Montague, 1974], permet d'associer à une analyse compositionnelle de la syntaxe de la phrase une interprétation dans un modèle donné de sa signification. En gardant les principes Montagoviens à cœur et en raffinant système de types, lexique et calcul, nous nous proposons d'intégrer à un canevas bien connu et utilisé par de nombreux outils une extension basée sur [Pustejovsky, 1995], en en implémentant les principes essentiels de ce dernier.

Dans le cadre des travaux de cette thèse, nous avons été aménés à proposer de nouveaux modèles de la composition dans [Mery et al., 2007a] et [Mery et al., 2007b], qui ont été élaborés dans un canevas complet dans [Bassac et al., 2010], auquel ont été ajoutés des principes de description lexicale dans [Mery, 2009]. Ce travail de synthèse et de complément sera exposé en détails dans le reste de ce manuscrit.

Chapitre 3

Modèles de la composition

3.1 État actuel de la composition

En se proposant d'asseoir notre système sur le calcul sémantique de Montague, nous ne souhaitons pas reprendre l'analyse syntaxique d'une phrase ou le choix d'un modèle pour son interprétation finale ; nous nous situons au niveau du calcul sémantique compositionnel qui, en l'état, ne prend pas en compte les informations lexicales riches nécessaires au traitement des phénomènes-cible.

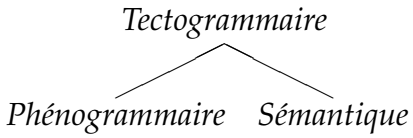
3.1.1 Une approche Montagovienne incrémentielle

Montague, dans [Montague, 1974], introduit le calcul logique qui se rendit célèbre sous le nom de "grammaire de Montague" : la Logique Intensionnelle. Un but motivait Richard Montague lors de sa construction : aboutir à une interprétation des phrases dans certains modèles, dont ceux de la théorie des ensembles, et ainsi furent définis les cadres sémantiques soutenus par cette logique qui n'avait qu'un but utilitaire à ses yeux. Cependant, le principe même de l'utilisation d'une logique pour représenter la sémantique d'une phrase, ainsi que la fondation de cette forme logique sur la structure syntaxique de la phrase, sont à l'origine de ce qui a été retenu de [Montague, 1974] par les spécialistes du traitement automatique des langues. Nous nous situons dans cette tradition marquée par le principe de compositionnalité : le sens d'une phrase, en tant que forme logique, est la composition du sens de ses parties.

L'analyse du sens d'une phrase passe donc par la combinaison, au moyen d'un calcul Montagovien, du sens de chacun des mots qui la composent.

3.1.2 Le cadre actuel des analyses sémantiques

Actuellement, et comme Chomsky le propose dès [Chomsky, 1957], l'approche de Montague est décomposée en deux étapes, tenant compte de la *structure profonde* d'une phrase : analyse syntaxique et analyse sémantique. Cette structure, héritée de Curry, est supposée être le parent commun de la sémantique (son interprétation dans un modèle) et de la syntaxe (sa réalisation dans le langage) pour une même phrase. [Muskens, 2010], entre autres, décrit le processus d'interprétation du langage comme l'arborescence suivante :



La Tectogrammaire est utilisée pour former de la structure profonde de la phrase. Elle décrit les applications des prédicats aux arguments.

La Phénogrammaire est utilisée pour décrire le langage lui-même, c'est-à-dire la structure de surface de la phrase : sa *syntaxe*.

La Sémantique est utilisée pour donner l'interprétation de la phrase : sa dénotation dans un modèle, tel celui des ensembles d'objets « réels ».

Les analyses de formalismes tels les ACG (décrites dans [de Groote, 2001]) se basent sur ce schéma pour proposer une formalisation en λ-termes de chaque niveau, et ainsi donner une analyse en deux étapes indépendantes d'un texte (dériver la structure profonde de la syntaxe, puis en déduire la sémantique), ou, de même façon, proposer une méthode de génération à partir de données sémantiques (en calculant la structure profonde de la donnée, puis en engendrant la syntaxe du texte voulu).

Dans un tel schéma, l'analyse de la sémantique lexicale fait partie de l'analyse de la sémantique de la structure profonde : supposons analysée la syntaxe d'une phrase et sa structure profonde calculée. Alors intervient le *lexique*, qui donne la signification de chaque terme. Avant de recomposer ces signifiants élémentaires suivant la structure calculée précédemment, il s'agit de prendre en compte les informations supplémentaires apportées par le lexique afin de traiter les cas détaillés par [Pustejovsky, 1995].

3.2 Types et ontologie

Nous avons choisi de rester près de [Pustejovsky, 1995] en adoptant une logique à n sortes, basée sur une ontologie de types, proche de TY_n.

3.2.1 Pourquoi un système de types ?

Le système de types, formellement, donne des contraintes sur les termes employés par la logique. S'il est commun, dans les approches Montagoviennes, d'utiliser les types pour exprimer le nombre d'arguments attendus pour un prédicat et leur ordre (sont-ce des objets élémentaires, des propriétés, des ensembles d'objets ?), [Pustejovsky, 1995] prend le parti d'utiliser également les types pour discriminer certaines propriétés attendues par les prédicats pour leurs arguments (s'agit-il d'un objet qui se mange, qui peut agir, etc. ?). De même, notre proposition est fondée sur les types ontologiques évoqués par le lexique génératif, et disposera également d'opérations sur ces types et d'opérations dirigées par ces types.

3.2.2 TY_n

Muskens détaille la logique TY_n dans [Muskens, 1996], et cette dernière est également utilisée par Hinderer dans la formalisation de [Hinderer, 2008]. Le principe est de remplacer la logique intentionelle de Montague proposée dans [Montague, 1974] par une logique à n sortes, c'est-à-dire comportant n types atomiques, définis dans le lexique, en plus du type t des valeurs de vérité. Cette logique a d'ores et déjà de nombreuses applications en elle-même, et c'est un canevas semblable qui est supposé par la hiérarchie ontologique de types du lexique génératif.

3.2.3 Hiérarchie ontologique de types

Le lexique génératif, comme vu précédemment, suppose une ontologie de types. En lieu et place du type ayant pour valeur les entités dans la logique intensionelle de Montague (e), nous disposons de types permettant de différencier chacun des phénomènes constatés. De plus, ils disposent d'une hiérarchie ontologique (donc d'une certaine notion de sous-typage, le type le plus général étant le e Montagovien). Au lieu de reprendre directement une hiérarchie existante ou de proposer la notre, nous supposerons son existence, et supposerons également qu'elle est construite comme une arborescence.

Chaque nœud (ou feuille) de la hiérarchie est une sorte de la logique, et cette dernière comporte donc autant de types, plus le type t des valeurs de vérité. Cette hiérarchisation nous permet, sans aucun autre artifice, de disposer d'un pouvoir d'expression plus grand que la logique intensionnelle simple, par exemple en ce qui concerne les quantificateurs généralisés.

3.2.4 Notion de sous-typage

Cette hiérarchie pourrait se modéliser suivant une notion de sous-typage, mais il nous semble plus intéressant d'utiliser des transformations systématiques pour représenter la relation d'héritage. En effet, le sous-typage est un mécanisme difficile à gérer de par lui-même, qui devient très problématique lorsqu'on l'utilise au second ordre.

3.2.5 Hiérarchie d'héritage

La notion de types utilisée ici est proche de celle du lexique génératif, qui n'a jamais été par elle-même définie entièrement. Cette notion peut être rapprochée des définitions et discussions formelles sur les structures de traits typés décrites dans [Carpenter, 1992]. Ainsi, la hiérarchie issue de l'ontologie constitue une hiérarchie d'héritage, définie comme un ordre partiel complet fini et borné sur l'ensemble des types et la relation de sous-typage ; ces types forment alors une structure de treillis.

[Carpenter, 1992] décrit également les notions formelles de typage, les mécanismes d'introduction de traits et d'inférence que l'on peut utiliser pour effectuer un raisonnement formel sur des structures comparables à celles du lexique génératif ou, par exemple, de HPSG.

3.2.6 Les relations

Nous disposons également de fonctions permettant le *changement de type*, qui modélisent les transformations induites par le lexique génératif. Ces fonctions transforment un objet d'un type A en objet d'un type B, et dessinent de cette façon des relations entre les types A et B ; on parlera de transformations ou de morphismes. Le graphe des relations entre types vient s'ajouter à la hiérarchie définie plus haut.

3.2.7 Illustration : quantification généralisée

Soit la locution *la plupart des étudiants ont réussi leur examen*. Une modélisation de cette phrase, (à l'interprétation très peu naturelle) en logique intensionnelle, calquée sur la quantification universelle, nous donne $(\textbf{plupart}\ x^{e}\ .\ (\text{etudiant}\ x) \rightarrow (\text{examen}\ x))$, et donc une quantification portant sur l'ensemble des entités ; une logique distinguant des types raffinés aurait donné la modélisation suivante : $(\textbf{plupart}\ x^{Etudiant}\ .\ (\text{examen}\ x))$, soit une quantification plus exacte et directe sur l'ensemble des étudiants. On distingue donc un quantificateur spécifique à chaque type.

Chapitre 4

Le comportement logique des phénomènes-cible

Avant de donner notre proposition de solution, détaillons ici un large échantillon des phénomènes posant problèmes : cas élémentaires de coercition, co-prédications élégantes ou non, problèmes de portée de quantification... Nous donnerons directement une ébauche logique correspondant à la solution recherchée, et nous intéresserons surtout aux cas limites.

4.1 Cas élémentaires

4.1.1 Polysémie contrastive

La « polysémie accidentelle » peut être traitée sans opérations lexicales (ou presque) ; il suffit de mettre en évidence la différence des types raffinés.

Ce bar a une bonne ambiance.

Pour simplifier, on considère le mot *Bar* comme ayant deux types raffinés possibles : celui de *Lieu*, correspondant au débit de boisson, et celui d'*Animal*, correspondant au poisson. Le prédicat *a une bonne ambiance* (abrégé *Ambiance*) porte sur les lieux.

$$\text{Syntaxe} : (\textit{Ambiance Bar})$$

$$\text{Type1} : (\lambda x^{Lieu}.(\text{Ambiance}^{Lieu \to t}\ x))\ \text{Bar}^{Lieu} \to (\text{Ambiance Bar})^{t}$$

$$\text{Type2} : (\lambda x^{Lieu}.(\text{Ambiance}^{Lieu \to t}\ x))\ \text{Bar}^{Animal} \to -$$

Un cas permet l'application directe, l'autre non.

Par contre, une première opération est utilisée implicitement pour les cas de sous-spécification, comme *voir* (les deux parties de la polysémie accidentelle peuvent être utilisées, car la construction de l'occurrence ne permet pas de les départager)

Je vois le bar.

$$\text{Syntax} : (\mathit{Voir}\ \mathit{Bar})$$

$$\text{Type1} : (\lambda x^{\varphi}.(\text{Bar}^{\varphi \to t}\ x))\ \text{Bar}^{Lieu} \to (\text{Voir}\ (f_{L\to\varphi}\text{Bar}))^{t}$$

$$\text{Type2} : (\lambda x^{\varphi}.(\text{Bar}^{\varphi \to t}\ x))\ \text{Bar}^{Bar} \to (\text{Voir}\ (f_{A\to\varphi}\text{Bar}))^{t}$$

Le prédicat requiert un attribut de la hiérarchie de *Physique* (φ), à laquelle appartiennent les deux termes considérés. Par conséquent, les deux peuvent convenir.

4.1.2 Qualia

Les cas « simples » d'exploitation de Qualia sont des opérations directes entre types : le télique d'une cigarette correspond à la consommation de celle-ci, et transforme donc un objet physique en événement, etc.

Formel

Les relations du quale « formel » sont ce que l'on a parfois appelé *accommodation* : des relations dans la hiérarchie des types. On considère que chaque objet dispose d'opérations canoniques vers ses ancêtres dans la hiérarchie, ce qui permet de traiter de manière uniforme l'hyponymie et les autres opérations.

Voiture lourde.

$$\text{Hierarchie} : \mathit{Voiture} \to \mathit{Vehicule} \to \mathit{Artefact} \to \mathit{Physique}$$

$$(\lambda x^{\varphi}.(\text{Lourd}^{\varphi \to t}\ x))\ \text{Voiture}^{Voiture} \to (\text{Lourd}\ (f_{Voiture\to\varphi}\ \text{Voiture}))^{t}$$

Nourriture chère.

$$\text{Hierarchie} : \mathit{Nourriture} \to \mathit{Bien}$$

$$(\lambda x^{Bien}.(\text{Cher}^{Bien \to t}\ x))\ \text{Nourriture}^{N} \to (\text{Cher}\ (f_{N\to Bien}\ \text{Nourriture}))^{t}$$

Film Réaliste.

Hierarchie : $Film \rightarrow Recit$

$$(\lambda x^{Recit}.(\text{Realiste}^{Recit \rightarrow t}\ x))\ \text{Film}^{Film} \rightarrow (\text{Realiste}\ (f_{F \rightarrow R}\ \text{Film}))^{t}$$

Dans les trois cas, il s'agit simplement de déterminer quel est le type ancêtre pertinent.

Constitutif

Le quale « constitutif » ne diffère que peu du formel. Il s'agit de propriétés saillantes du terme par relations méronymiques. Le lexique doit définir les opérations correspondantes, et ce seulement pour les relations usuelles – on doit pouvoir inférer d'autres relations, mais avec difficulté. Notons qu'il ne s'agit pas de relations entre *types* (bien que les types puissent servir de discriminant), mais entre *termes* lexicaux.

Ordinateur rapide.

Relation : $CPU \in Ordinateur$

$$(\lambda x^{P^*}.(\text{Rapide}\ x))\ \text{Ordinateur}^{Artifact} \rightarrow (\text{Rapide}\ (f_C\ \text{Rapide})$$

(Remarque : Le prédicat *rapide* accepte un grand nombre de types, par analogie avec un objet physique se déplaçant à grande vitesse. Il s'agit d'un sous-ensemble de φ, difficilement définissable avec précision, noté ici P^*.)

Voiture économe.

Relation : $Moteur \in Voiture$

$$(\lambda x^{M}.(\text{Econome}\ x))\ \text{Voiture}^{Voiture} \rightarrow (\text{Econome}\ (f_C\ \text{Voiture})$$

(Remarque : On pourrait généraliser avec une opération portant sur un type complet tel que *Véhicule*, mais cela semble hasardeux. On suppose systématiquement que les voitures fonctionnent avec un moteur à explosion, ce qui correspond à la situation courante du locuteur – mais qui est un exemple d'opération dépendant assez fortement de l'univers d'énonciation (cette phrase n'a pas de sens dans un autre contexte que contemporain).)

Pomme rouge.

Ici, l'exemple est beaucoup moins tranché. On fait clairement référence à la pigmentation du tégument, mais il paraît difficile d'affirmer qu'on effectue une exploitation. Dans l'hypothèse de simplification maximale, on aurait une prédication directe (les prédicats chromatiques s'appliquent à tout objet physique, et donc aux pommes). Dans l'hypothèse stricte (seule la partie visible / extérieure doit être sélectionnée), le type visé est assez spécifique ; notons-le *Vis*, et...

$$\text{Relation : } Peau^{Vis} \in Pomme^{Nourriture}$$

$$(\lambda x^{Vis}.(\text{Rouge } x))\ \text{Pomme}^{Nourriture} \rightarrow (\text{Rouge } (f_C\ \text{Pomme})$$

(Remarque : On utilise f_C pour « accès au quale constitutif », mais il peut y avoir plusieurs possibilités à ce stade. Dans *Une voiture sale*, les parties sales peuvent être : vitres, carrosserie, intérieur... sans exclusivité (le type *Vis* est ici très pertinent).)

Agentif

Le quale « agentif » fonctionne toujours de la même manière. En général, cependant, il fait référence à un même type – celui des agents (A). On notera f_A l'opération d'accès à l'auteur, le réalisateur... etc, d'un objet ; Pustejovsky note que les termes pouvant être concernés appartiennent à un sous-type de *Artefact*, soit dans la hiérarchie physique, soit dans la hiérarchie abstraite. Le lexique doit préciser, s'il y a lieu, cette relation. *Important* : il peut très facilement y avoir une ambiguïté dans ces constructions, voir dans les exemples.

Un article partial

$$\text{Relation 1 :} Author(Article, Journaliste)\ Journaliste \in A$$

$$\text{Relation 2 :} Author(Article, Editorialiste)\ Editorialiste \in A$$

$$\text{Relation 3 :} Author(Article, Corporation)\ Corporation \in A$$

$$(\lambda x^{A}.(\text{partial}^{A \rightarrow t}\ x)\ \text{Article} \rightarrow$$

$$(\text{partial}\ (f_A\ \text{Article})^{A})$$

(Remarque : la valeur de f_A est sous-spécifiée au moins entre les trois relations citées ici. Le type ne permet pas de faire la distinction. Objectivement, la phrase non plus. Attention : on doit tout de même n'inclure lexicalement que les agents possibles pour l'article en tant qu'information ; la relation $Author(Article, Printer)$ ne doit pas apparaître...)

Un design intéressant

$$\text{Relation} : Author(A, Design)$$
$$(\lambda x^{A}.(\text{interessant}^{A\to t}\ x)\ \text{Design} \to$$
$$(\text{interessant}\ (f_A\ \text{Design})^{A})$$

(Remarque : ici, *Design* est simplement un objet abstrait et est, en lui-même, très ambigu. La seule garantie pour l'agentif est qu'il existe au moins un agent qui en soit à l'origine, mais cette relation est *fortement* sous-spécifiée.)

Une pièce inspirée

$$\text{Relation 1} : Author(Piece, Dramaturge)\ Dramaturge \in A$$
$$\text{Relation 2} : Author(Piece, Acteur)\ Acteur \in A$$
$$(\lambda x^{A}.(\text{inspire}^{A\to t}\ x)\ \text{Piece} \to$$
$$(\text{inspire}\ (f_A\ \text{Piece})^{A})$$

(Remarque : ici, la polysémie porte également sur l'argument. La pièce de théâtre est-elle envisagée en tant qu'œuvre littéraire, ou en tant que spectacle vivant ? De quelle *facette* parle-t-on ? La phrase ne permet pas de le déterminer. Il s'agit là d'un prototype de *Dot object*.)

Télique

Le quale « télique » complète la structure de qualia. Il s'agit d'identifier un comportement typiquement associé (par destination ou par dessein) ; en général, ce quale fait référence à un terme de type *événement* (ou d'un type en descendant). Un même objet peut éventuellement avoir plusieurs téliques. La relation est entre termes, et est lexicalisée.

Une cigarette relaxante

$$\text{Usage} : (\text{fumer cigarette})^{Evt}$$
$$(\lambda x^{Evt}.(\text{relaxant}\ x)\ \text{Cigarette}\ \to\ (\text{relaxant}\ (f_T\ Cigarette))$$

(Remarque : ici, la difficulté est dans la surgénération : peut-on vraiment conclure que le locuteur aime utiliser l'objet pour son télique, ou apprécie simplement sa possession ? Seul le contexte peut déterminer l'interprétation correcte. La vision de Pustejovsky (priorité au télique) est donnée ici, mais ce n'est pas dogmatique.)

Un verre fort

Usage :$(\text{contenu verre})^{Boisson}$

$(\lambda x^{Boisson}.(\text{fort } x)\ \text{Verre} \rightarrow (\text{fort } (f_T\ Verre))$

(Remarque : il s'agit de l'analyse la plus correcte pour conteneur / contenu. On peut tenter une analyse de type *Packing* ou *Dot*, mais le résultat est bien plus spécieux. Le prédicat *fort* ne vise pas un événement, mais une classe également assez sous-spécifiée comprenant *Boisson*, *Nourriture*, etc.)

Une bonne pierre

(Remarque : nous sommes ici dans un cas classique de sous-spécification pathologique où on ne sait absolument pas, en-dehors d'un contexte supplémentaire, à quoi est bon le caillou en question – il s'agit toujours d'un événement. L'analyse est la même, mais le $f_T^{\varphi \rightarrow Evt}$ est complètement inconnu ; on sait simplement qu'il existe. Le locuteur n'a pas de problème avec ce genre de phrases ; il attend simplement une explication.)

Transformations lexicales

Grinding et *Packing* sont des opérations lexicales semblables à l'exploitation du télique, mais différentes en ce qu'elles sont localement destructrices (il est impossible pour un lexème d'être référé simultanément par deux états différents quand ces derniers sont liés par une de ces relations). Au début, ces relations étaient expliquées par des *Dot*, ce qui posait les problèmes de la co-prédication...

Les exemples se ressemblent tous ; il est raisonnable de penser que les opérations (cuisine, tissage) soient génériques, et se décrètent donc au niveau des types. On les note, pour *Grinding*, f_G, elles sont peu nombreuses.

Animal/Nourriture, Animal/Artefact
Également réalisable avec végétaux : *citronelle fraîche*, *lin blanchi*, etc.

Un saumon délicieux

Operation : cuisine —$Animal \rightarrow Nourriture$

$(\lambda x^{Nourriture}.(\text{delicieux } x)\ \text{Saumon}^{Animal} \rightarrow (\text{delicieux } (f_G\ Saumon))$

Materiau/Liquide
Cet exemple est restreint à certaines catégories spécifiques. Il est aussi plus convaincant en Anglais.

Café tiède

$$\text{Operation : liquefaction} \text{—}\{\text{cafe, the, menthe, teinture}\ldots\}^{Materiau \to Liquide}$$

$$(\lambda x^{Liquide}.(\text{tiede } x)\ \text{Cafe}^{Plante} \ \to\ (\text{tiede } (f_G\ Cafe))$$

4.1.3 Résultatifs

[Jacquey, 2001] est fondée sur le traitement de ces cas particuliers. Si [Pustejovsky, 1995] et les travaux subséquents tendent à les assimiler aux •-objets, nous pensons que leur cas peut être traité à part. Il s'agit des éléments lexicaux comme *rédaction*, *construction*… qui ont pour aspects le processus élaboratif et le résultat de ce processus. Par exemple :

La rédaction est en cours : processus

La rédaction est à rendre avant lundi : résultat

4.1.4 Transferts

Les exemples classiques de [Nunberg, 1993] sont également une catégorie d'opérations lexicales, dépendant ici du prédicat et non pas de l'argument. Citons notamment :

Je suis garé en double-file : *je* en tant que véhicule.

La table a commandé un churrasco : *la table* en tant que personnes.

4.1.5 Facettes

Les types pointés, *Dot Objects* ou lexèmes multifacettes sont une partie importante du lexique génératif. Il s'agirait d'objets « complexes » en ce qu'ils semblent disposer de deux types (voire plus). Il est possible d'adhérer ou de nier cette analyse ; voyons ici l'ensemble des cas, sans proposer d'analyse logique à ce stade. L'exemple canonique est *livre* : on peut en effet parler d'*un livre épais mais très intéressant*, ce qui reflète bien deux aspects apparemment incompatibles pour ce terme…On distingue les *Dots* « lâches » et « liés » :

Facettes lâches

Les *Dots* « lâches », dans notre terminologie, correspondent à des aspects multiples d'un même lexème qui sont circonstanciellement liés, mais pour lesquels il est difficile de fixer un type : ainsi Pustejovsky affirme-t-il qu'un repas est composé de l'événement et de la nourriture, mais on pourrait également y associer d'autres aspects, comme les convives. Le type lexical le plus prolixe dans ses aspects est celui des villes, qui peut présenter un nombre impressionnant de facettes et est très facilement productif.

Événement •Information

Cours, conférence, briefing, réunion, colloque, allocution… L'événement dispose d'un aspect informatif fortement saillant. Autres aspects accessibles : l'allocutaire (via l'agentif), le lieu.

Un cours passionnant.

Un cours qui dure.

Événement•Musique

Concert, récital. Autres aspects accessibles : les concertistes, les instruments, le lieu, le public.

Un concert de Jazz.

Le concert de Noël.

Spectacle•Musique

Voir *Événement•Musique* (la différence est qu'il ne peut s'agir d'un enregistrement). On dispose également de *Spectacle•Information* pour les pièces de théâtre, à rapprocher de *Événement•Information*.

Événement•Physique

Déjeuner, dîner, toast. Autres aspects accessibles : convives, service, lieu, élément du mobilier (buffet…).

Un fort bon repas

Un repas à l'heure

Organisation.—

Journal, école, temple… L'aspect dominant est celui de l'organisation, auquel peuvent s'ajouter des publications, des communiqués, divers activités, et très souvent des groupes de personnes et un ou des lieux.

L'école est au bout de la rue.

L'école est en grève

L'école est fermée

Lieu.Personnes.—

Les villes. Les aspects sont fortement divers.

Paris n'est plus une ville aussi vivante qu'autrefois.

Paris est relativement petite.

Paris devient de plus en plus écologiste.

Paris essaye désespérément de vendre des avions au Brésil.

Paris n'a aucune chance pour ce championnat.

Paris est source constate d'inspiration et d'émerveillement.

Je ne supporte pas Paris.

Facettes liées

Les *Dots* liés sont, *a contrario*, des lexèmes aux aspects étroitement dépendants l'un de l'autre. Ils sont moins nombreux et plus déterminés, et les multiples prédications semblent plus faciles, naturelles, ou élégantes. L'objet emblématique de cette classe est *Livre*, qui sert d'exemple presque canonique.

Physique•Ouverture

Portes et fenêtres. Le phénomène est assez restreint.

Un portail de fer forgé.

Un portail ouvert.

Acte•Proposition

Questions et réponses. Le langage identifie presque les deux aspects, mais la logique les distingue (on peut poser bien des fois une même question).

Je n'ai pas entendu votre question.

Cette question est intéressante.

État •Propostion

Opinions, croyances…

Ma foi est forte.

Je crois en un macrocosme absolu.

Attribut •Valeur

Toute valeur numérique associée à une variable.

Les niveaux de CO_2 sont à $385ppm$.

Rien n'a été fait pour empêcher les niveaux de CO_2 d'augmenter.

Information•Physical

Les livres, et de nombreuses versions de supports écrits de l'information (manuscrit, rouleau, affiche, tablette, etc.). L'exemple canonique.

Un livre intéressant.

Un livre lourd.

Sound.Information.—

Analogue au précédent, sauf qu'on y ajoute un support sonore, pour tout enregistrement. Le troisième aspect est physique, informatique... On peut également ajouter, sur le même modèle, l'ensemble des supports multimedia.

Un CD gravé à la maison.

Un CD de Jazz.

4.2 Forçage de l'aspect

En plus des constructions de la langue permettant l'accès aux divers aspects d'un terme, il peut également être fait référence aux aspects de façon explicite, au moyen de locutions comme *en tant que* ou *considéré comme*.

[Asher, 2011] consacre une attention spécifique à ces constructions. Citons notamment :

En tant que banquier, Jean gagnait 500 000 euros par an.

En tant que balayeur, Jean ne gagne plus que 15 000 euros par an.

4.3 Co-prédications

Le test de la *co-prédication* est un élément majeur de [Pustejovsky and Asher, 2000] et des travaux qui ont suivi, et est effectivement un bon moyen de caractériser la compatibilité d'opérations portant sur les divers aspects d'un même terme. Le principe est d'appliquer deux prédicats sélectionnant deux aspects différents à un même terme, et de faire examiner par le locuteur le résultat ; l'exemple de *un livre lourd mais intéressant* a fortement influencé la théorie des •-objets. Examinons.

4.3.1 Co-prédications élégantes

Un grand nombre de co-prédications sont *élégantes*, et les divers aspects examinés sont facilement compatibles, même eu égard à leur différence conceptuelle a priori. Voici quelques exemples :

Qualia

Une voiture rouge puissante (Formel, Constitutif)

Un journal engagé et peu cher (Agentif, Formel)

Un article neutre et agréable (Agentif, Télique)

Facettes

Un livre lourd et intéressant (Physique, Information)

La porte bleue est ouverte (Physique, Ouverture)

Une température de trente degrés allant en augmentant (Valeur, Attribut)

Paris, née sur les rives de la Seine, est une ville de cadres dynamiques au patrimoine sans pareil, qui souhaite se développer plus avant (Ville, Lieu, Population, Bâtiment, Institution)

Qualia et facettes

J'ai commencé le livre sans couverture (Télique, Physique)

4.3.2 Co-prédications douteuses

À l'inverse, certaines co-prédications plus douteuses viennent mettre en lumière certains processus incompatibles avec d'autres

Grinding

Ce saumon était rapide et délicieux. (?)

Ici, on a une claire contradiction entre l'aspect avant et après l'opération destructive. Cependant, il est acceptable de dire :

Ce saumon, qui était rapide, est délicieux.

Et également :

Ce saumon était rapide. Il est délicieux.

Ce qui indique une certaine flexibilité vis-à-vis des reprises anaphoriques de l'argument, pour le moins.

Facettes

Paris, célèbre pour ses monuments Napoléoniens et ses habitants pressés, a refusé d'attaquer l'Irak. (?)

Ici, il semble difficile d'utiliser un aspect en particulier (celui de gouvernement d'un pays) en conjonction avec les autres aspects normalement accessibles. Il est également difficile d'y faire référence en utilisant les constructions anaphoriques vues précédemment :

Paris, qui est célèbre pour ses monuments Napoléoniens et ses habitants pressés, a refusé d'attaquer l'Irak. (?)

Paris est célèbre pour ses monuments Napoléoniens et ses habitants pressés. Elle a refusé d'attaquer l'Irak. (?)

4.3.3 Conclusion

On peut distinguer de multiples classifications pour les phénomènes de sémantique lexicale, mais trois classes se dégagent opérationnellement. Les aspects accessibles « normalement », qui disposent d'une grande flexibilité dans leurs utilisation. Les aspects correspondant aux opérations destructrices, qui résultent en des co-prédications douteuses, mais restent co-accessibles avec les autres aspects sous certaines constructions anaphoriques. Et enfin les aspects telles le *gouvernement* du pays dont une ville est capitale, très particuliers, qui résistent à la co-prédication même en cas de reprise anaphorique du terme, et sont donc très rigides dans leur utilisation.

C'est cette échelle de flexibilité / rigidité qui nous intéressera, bien plus que la différence entre *qualia* et facettes, dans notre formalisation. Examinons les modifications induites par la syntaxe sur cette rigidité :

4.3.4 Conséquences de la syntaxe des phrases complexes

Le problème est le suivant : dans une phrase complexe (plus complexe qu'une simple prédication), comment la syntaxe dirige-t-elle les opérations lexicales pour les composantes internes de l'argument et du prédicat ? Quel est l'impact de la flexibilité (le caractère local/global) des opérations de ce point de vue ? Rappelons notre argument central : la sémantique est *compositionnelle*, et le sens d'une phrase dépend de la combinaison, guidée par la syntaxe et la structure tectogrammatique induite, du sens des mots qui la composent.

Argument complexe

Reprenons quelques exemples basiques :

Un article partisan

Un poulet savoureux

Un livre lourd mais intéressant

Paris a refusé d'attaquer l'Irak

Il semble assez clair que la complexité syntaxique de l'argument n'ait pas d'impact réel sur l'opération effectuée, on peut se contenter du nominal sémantique saillant résultant de toutes les combinaisons possibles pour retrouver les mêmes données.

Un court article partisan de Fodor & Lepore

Un poulet massamba, préparé traditionnellement, savoureux

Une antologie de science-fiction des meilleurs auteurs des années 1980 lourde mais intéressante

Paris, malgré un gouvernement marqué par une alliance traditionnellement forte avec Washington, a refusé d'attaquer l'Irak

Une fois que la syntaxe a identifié les arguments des prédicats *partisan*, *savoureux*, *lourd mais intéressant*, *a refusé d'attaquer l'Irak* à savoir des versions fortement modifiées certes, mais ramenant tout de même à *article*, *poulet*, *antologie* et *Paris*, les processus sont exactement les mêmes que dans le cas de la prédication simple (avec un sous-typage en plus pour *antologie*).

Prédication complexe

Avec les mêmes exemples de base, nous pouvons réaliser des prédications plus complexes – qui seront alors des cas de co-prédication. L'important est de distinguer la syntaxe utilisée pour ces prédicats.

Prédicats coordonnés

Un article partisan et long

Un poulet savoureux et vif (?)

Un livre lourd et intéressant

Paris a refusé d'attaquer l'Irak et se situe au Nord de la France (?)

Nous avons ici quatre exemples "classiques" de co-prédication, seul le second et le quatrième sont mauvais de par la "rigidité" déjà évoquée des constructions associées.

Une prédication, un objet

Cet article partisan est très long

Ce poulet savoureux est très vif (?)

Ce livre lourd est très intéressant

Paris, au Nord de la France, a refusé d'attaquer l'Irak (?)

Les phénomènes se comportent strictement de la même façon que précédemment

Une relative

Cet article, qui est partisan, est très long

Ce poulet, qui était très vif, est savoureux

Ce livre, qui est intéressant, est très lourd

Paris,qui se situe au Nord de la France, a refusé d'attaquer l'Irak (?)

Ici, on remarque une acceptibilité plus grande pour une co-prédication avec *grinding* que précédemment. Le verrouillage de la "Capitale de la France" reste, lui, présent.

Une reprise anaphorique dans une phrase postérieure

Un article partisan. Il est également assez long.

Un poulet savoureux. Il était très vif.

Un livre lourd. Il est très intéressant.

Paris a refusé d'attaquer l'Irak. Elle se situe au Nord de la France. (?)

On observe ici le même comportement qu'avec les relatives.

Phrase complexe

Une analyse peu poussée permet de comprendre facilement que les mêmes phénomènes que les prédications complexes se produisent en mêlant prédications et arguments complexes.

Conclusions préliminaires

Sans s'avancer beaucoup, on peut faire les trois observations suivantes et penser à les généraliser :

1. Quelle que soit la complexité de l'argument, les mêmes opérations lexicales s'y appliquent.
2. Une co-prédication est toujours possible pour une certaine catégorie d'opérations lexicales, dites flexibles ; pour d'autres opérations, dites semi-flexibles seule une forme de référence par pronom ou anaphore permet la co-prédication. Ces classes sont stables sur l'ensemble des structures syntaxiques étudiées.
3. Une dernière classe d'opérations est dite rigide, et ne permet en aucun cas la co-prédication, quelle que soit la structure syntaxique employée.

4.4 Quantifications

Dans ses travaux récents, Asher met en lumière un phénomène intéressant, celui de la quantification sur différents aspects d'un même terme, dans des phrases comme :

J'ai lu tous les livres de la bibliothèque.
(Le locuteur a lu à chaque fois une fois le contenu informationnel de tous les livres physiquement présents, en sautant présumablement les copies en double exemplaire.)

J'ai volé tous les livres de la bibliothèque.
(Ici, il s'agit probablement de l'ensemble des copies physiques.)

Tous les livres de la bibliothèque ont brûlé. Heureusement, je les avais déjà lus.
(La sélection porte d'abord sur l'ensemble physique, puis sur un sous-ensemble informationnel.)

Ces exemples prouvent qu'un mécanisme de quantification générique ne peut suffire à exprimer correctement l'ensemble des situations décrites ; la quantification, et l'expression des différents aspects d'un même terme, doivent faire preuve d'un certain raffinement dans leur formalisation.

4.5 Mécanismes extraphrasiques

Dans ces mêmes travaux (en lien avec ses préoccupations liées au discours exprimées, par exemple, dans [Busquets et al., 2001]), Asher signale que certains termes peuvent changer d'aspect (notamment de télique) dans un certain contexte discursif, par une composition extérieure à la phrase. Ainsi, le texte suivant :

J'ai commencé par la cuisine, puis je me suis attaqué à la pièce principale avant d'avancer dans les chambres.

prend un sens tout différend selon qu'on le précède de :

Hier, j'ai nettoyé la maison.

ou de :

Hier, j'ai mené l'assaut sur les terroristes.

Deuxième partie

Solution proposée

Chapitre 5

Mécanismes d'ordre supérieur

Notre proposition est fondée sur un calcul logique incluant de multiples types atomiques et un liage au second ordre.

5.1 ΛTY_n

Nous nous basons sur la logique typée à n sortes (TY_n) en étendant le calcul au second ordre (à la manière du Système F de Girard, voir [Girard, 1972]). Nous nommons ΛTY_n le calcul résultant, et nous allons détailler dans cette partie les formules de la logique, les termes du calcul et leur relation. (La preuve de certaines des propriétés de ce calcul a fait l'objet de travaux approfondis, [Retoré, 2011], qui seront développés et publiés ultérieurement.)

5.1.1 Propriétés élémentaires

Lexique

Voici l'ensemble des termes élémentaires utilisés dans le calcul.

- Soit $P = \{a, b, c, d, e, \ldots\}$ un ensemble de types. On parle de logique à n sortes quand $|P| = n$, et on note cette logique TY_n. Soit également t un type, $P \cap \{t\} = \varnothing$. Dans le cadre du λ-calcul du second ordre, on notera la logique correspondante ΛTY_n.

- Soit $C = \{A, B, \ldots\}$ l'ensemble des *constantes* du lexique. Ils ont, par définition, un *type associé* $type(A) \in P$ pour $A \in C$. On a également $\bot, \top \in C$ deux constantes, $type(\bot) = type(\top) = t$

- Soit $Q = \{R, S, \ldots\}$ l'ensemble des *prédicats* du lexique. Ils ont, par définition, un *type associé* $type(R)$ de la forme $\alpha_1 \ldots \alpha_k \rightarrow t$, avec les α_i des types de P et k l'*arité* du prédicat.

- Soit $F = \{f, g, \ldots\}$ l'ensemble des *fonctions* du lexique. Elles ont, par définition, un *type associé* $type(f) = \alpha_1 \ldots \alpha_k \rightarrow \beta$, où les α_i et β sont des types de constantes ou de fonctions, et k est l'arité de la fonction.

- Soient encore les prédicats suivants :
  - $\wedge$ (et), de type $t \rightarrow t \rightarrow t$.
  - $=$ (égal), de type $t \rightarrow t \rightarrow t$.
  - $\neg$ (négation), de type $t \rightarrow t$.
  - Pour chaque $\alpha \in P$, soit $\forall_\alpha$, de type $\alpha \rightarrow t \rightarrow t$.

- Soit le lexique, $L = C \cup Q \cup \{\wedge\} \cup \{\neg\} \cup \{=\} \cup \{\forall_\alpha | \alpha \in P\}$.

Nous avons donc défini ici : l'ensemble des types de la hiérarchie (les n sortes de la logique), les constantes et prédicats lexicalisés, les transformations, et les opérations permettant de reconstruire l'ensemble des opérateurs booléens : négation, conjonction, disjonction, implication, équivalence, égalité, différence. Nous introduisons également un quantificateur universel pour chacun des types atomiques, ce qui rejoint la volonté de quantification généralisée ; ils permettent de définir un quantificateur existenciel pour chaque type, symétriquement.

Termes

Nous préciserons ici les *termes* du λ-calcul typé au second ordre.

- Soit V un ensemble de variables. Toute variable $v \in V$ (de type arbitraire α), ou constante $c \in C$, (de type donné par le lexique) sont des termes.

- Pour τ un terme de type $\alpha \rightarrow \beta$ et ρ un terme de type α, on a $(\tau\ \rho)$ un terme de type β.

- De même pour une fonction f d'arité $k > 1$ et de type $\alpha_1 \ldots \alpha_k \rightarrow \beta$ et ρ un terme de type α_1, on a $(f\ \rho)$ un terme de type $\alpha_2 \ldots \alpha_k \rightarrow \beta$ qui sera une fonction d'arité $k-1$.

- Pour v une variable de type α et τ un terme de type β, on a $\lambda v^\alpha.\ \tau$ un terme de type $\alpha \rightarrow \beta$.

- Pour τ un terme de type $\alpha \to t$ et v une variable de type α, $\forall_\alpha v\ .\ \tau$ est un terme de type t..
- Soit $X = \{Y, Z \ldots\}$ un ensemble de *variables de types*. Pour $Y \in X$ et τ un terme de type α, on a $\Lambda Y.\tau$ un terme de type $Y \to \alpha$.
- Pour $Y \in X$ et pour τ un terme de type $Y \to \alpha$, $\tau\{Y\}$ est un terme de type α.
- Il n'y a pas d'autres termes.

Formules

Nous détaillons ici l'ensemble des formules de la logique concernée.

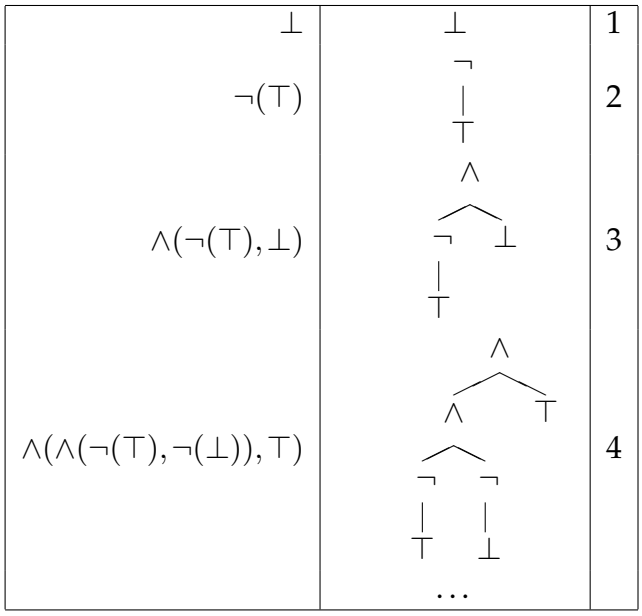
- $\top$ et $\bot$ sont des formules.
- Pour un prédicat R d'arité k et de type $\alpha_1 \ldots \alpha_k \to t$, pour des termes τ_i de type α_i $(1 \leq i \leq k)$, $R(\tau_1, \ldots, \tau_k)$ est une formule.
- Soit φ une formule. Alors $\neg\varphi$ est une formule, et (pour une variable x de type α) $\forall_\alpha x.\varphi$ est une formule.
- Soient φ et ψ deux formules. Alors $\wedge(\varphi\ \psi)$, $=(\varphi\ \psi)$ sont des formules.
- Il n'y a pas d'autres formules.

Propriétés des termes

- Un terme est *clos* quand il ne comporte aucune variable liée ($\lambda x.\tau$).
- Forme normale longue $\beta\eta$ (introduite dans [Huet, 1976]) :
  - Un terme atomique dont le type est soit un type de base, soit une variable de type, est sous forme normale longue $\beta\eta$.
  - Si τ est sous forme normale longue $\beta\eta$, alors $\lambda x.\tau$ et $\Lambda\alpha.\tau$ le sont aussi.
- D'après la propriété de Church-Rosser, un terme est *normalisable* quand il peut se réduire en une forme normale ; cette forme normale est unique (la propriété de Church-Rosser revient à la confluence pour les termes normalisables).

Taille d'un terme

- On écrit $||\tau||$ la taille (ou hauteur) de τ.
- $||v|| = ||c|| = 1$, pour $v \in V$ et $c \in C$.
- Pour tout autre terme τ, la taille du terme est égale à la hauteur de son arbre syntaxique :



- Le *degré de quantification* au premier ordre d'un terme est le nombre de variables liées à ce terme : k pour $\lambda x_1 \ldots x_i \,.\, \tau$.
- Le *degré de quantification* au second ordre d'un terme est le nombre de variables *de type* liées à ce terme : k pour $\Lambda \xi_1 \ldots \xi_k \,.\, \tau$.

Taille d'une formule

- $||\top|| = ||\bot|| = 1$
- $||R(\tau_1, \tau_2, \ldots, \tau_k)|| = 2$
- $||\neg\varphi|| = ||\forall_\alpha x\ \varphi|| = ||\varphi|| + 1$
- $||\wedge(\varphi\ \psi)|| = ||\{=\}(\varphi\ \psi)|| = max(||\varphi||, ||\psi||) + 1$

Propriété : toute formule est un terme de type t

- $\top$, $\bot$ sont deux termes de type t.
- Pour un prédicat R d'arité k et de type $\alpha_1 \dots \alpha_k \to t$, pour des termes τ_i de type α_i $(1 \leq i \leq k)$, $R(\tau_1, \dots, \tau_k)$ est un terme de type t.
- On procéde ensuite par induction sur la taille des termes.
- Soit φ un terme de type t. Alors $\neg\varphi$ est un terme de type t ($\neg$ étant de type $t \to t$), et (pour une variable x de type α) $\forall_\alpha x.\varphi$ est un terme de type t ($\forall_\alpha$ étant un terme de type $\alpha \to t \to t$).
- Soient φ et ψ deux termes de type t. Alors $\wedge(\varphi\ \psi)$, $=(\varphi\ \psi)$ sont des termes de type t, $\wedge$ et $=$ étant tous deux de type $t \to t \to t$.
- On a vu tous les cas possibles de formules : toutes sont donc des termes de type t.

5.1.2 Correspondance des formules et termes avec les λ-termes

Nous cherchons à démontrer que tout terme de type t, qui soit clos et normal, soit une formule.

Cas de base

Soit un terme de type t clos, normal et de taille 1. D'après la définition du langage, ce terme est forcément égal à $\top$ ou $\bot$, donc ce terme est une formule.

- Un terme τ *du premier ordre* de degré i sous forme normale s'écrit $\lambda x_1 \dots x_k\ .\ h\ \tau_1 \dots \tau_l$, avec h une variable et les τ_j des termes normaux et clos de degré maximum $i-1$. Hypothèse d'induction : si l'un des τ_j est de type t, alors c'est une formule ; s'il est de type α avec α un type atomique de P, alors c'est un terme de type α.
- Si τ est, de plus, de type dans $P \cup t$, alors $\tau = h\ \tau_1 \dots \tau_l$.
- Examinons les différentes valeurs possibles de h. Si on suppose τ *clos*, h n'est pas une variable. Il s'agit donc d'un terme d'arité l...
- Si $l = 0$, alors h est une constante : il s'agit soit d'un terme de type inclus dans P, soit de $\top$ ou $\bot$, de type t. Si τ est de type t, alors il s'agit d'un de ces derniers cas, et τ est une formule.

- Si $l = 1$, alors h est de type $\alpha \to \beta$, et $\tau = h^{\alpha\to\beta}\,\tau_1^\alpha$ est de type β. Si τ est de type t, alors :
  - Soit h est un prédicat R d'arité 1, auquel cas τ est une formule.
  - Soit $h = \neg$, auquel cas τ est une formule.
  - Il n'y a pas d'autre cas.
- Si $l = 2$ et que h est de type $\alpha \to \beta \to \delta$, alors $\tau = h^{\alpha\to\beta\to\delta}\ \tau_1^\alpha\ \tau_2^\beta$ est de type δ. Alors s'il s'agit de t :
  - Soit h est un prédicat R d'arité 2, auquel cas τ est une formule, car τ_1 et τ_2 sont des termes respectivement de types α et β de P, ou bien des formules.
  - Soit $h = \{=\}$, auquel cas τ_1 et τ_2 sont de type t ($=$ étant de type $t \to t \to t$), et τ est une formule.
  - Soit $h = \wedge$, auquel cas τ_1 et τ_2 sont de type t ($\wedge$ étant de type $t \to t \to t$), et τ est une formule.
  - Soit $h = \forall_\alpha$, auquel cas τ_1 est une variable de type α (x^α), τ_2 est de type t (φ^t), et $\tau = \forall_\alpha x\varphi$ est une formule.
  - il n'y a pas d'autre cas.
- Quand $l \geq 3$, h est de type $\alpha_1\alpha_2\ldots\alpha_l \to \beta$, et $\tau = h^{\alpha_1\alpha_2\ldots\alpha_l\to\beta}\ \tau_1^{\alpha_1}\ldots\ \tau_l^{\alpha_l}$ est de type β. Alors s'il s'agit de t :
  - Soit h est un prédicat R d'arité l, auquel cas τ est une formule.
  - Il n'y a pas d'autre cas.

5.1.3 Termes du second ordre

Formes des termes au second ordre

Un terme du second ordre est constitué d'un certain nombre (éventuellement nul) de λ et Λ-abstractions, et d'un sous-terme τ appliqué à des arguments qui sont des λ-termes (correspondant aux λ-abstractions, si elles existent) ou des types (correspondant aux Λ-abstractions, si elles existent).

Si le terme est normal : τ ne peut être une λ ou Λ-abstraction.

Nous pouvons considérer que les λ-termes normaux bien typés sont η-étendus, et que donc chaque terme de type $A \to B$ sans argument est de la forme $\lambda x^A.u^B$, de même que la forme d'un terme de type $\Lambda\alpha.\varphi$ est $\Lambda\alpha.u$.

Tout λ-terme normal de type t est une formule

Les mêmes arguments qu'au premier ordre s'appliquent. Dans le cas du second ordre, il ne peut y avoir de Λ-abstraction dans un terme normal de type t où les seules variables libres ont leurs types dans P.

5.1.4 Conclusion

Nous avons une correspondance complète entre termes normaux de type t et formules dans le calcul associé à ΛTY_n pour tout n.

5.2 Une modification de l'application

L'application prend effet comme d'habitude en cas d'accord de types. Si les types diffèrent, cependant, la logique du second ordre est utilisée pour anticiper la transformation d'un type encore inconnu vers un type apparaissant dans le terme.

Les éléments du lexique sont donnés avec des termes transformationnels qui modélisent des morphismes spécifiques pour résoudre ces différences de types. Il y a plusieurs façons de résoudre une transformation, selon le degré de flexibilité δ de cette dernière : si cette flexibilité est suffisante ($\delta(f) = 1$), on peut employer une transformation locale (2), sinon, on devra utiliser une transformation globale (1).

Soit la situation d'application problématique :

$$(\lambda x^V.\ (P^{V\to W} x))\tau^U$$

La divergence de types peut être résolue selon soit (1) soit (2), selon les termes disponibles :

1. Utilisation globale d'une transformation :

$$(\lambda x^V.\ (P^{V\to W} x))\ (f^{U\to V}\tau^U)$$

   Supposons que f est un terme optionnel associé soit avec P, soit avec τ. La transformation est appliquée directement à l'argument, quelle que soit la structure du prédicat – il peut y avoir de multiples occurrences de la variable x, et toutes résulteraient en une erreur de typage si la transformation f était indisponible. Par exemple, une conjonction serait résolue en $(\lambda x^V.\ (\wedge\ (P^{V\to W} x)\ (Q^{V\to W} x))\ (f^{U\to V}\tau^U)$, afin que chaque occurrence de l'argument soit transformée de la même façon.

Ici, le type de l'argument – et, par conséquent, le domaine de la transformation – est connu. Nous pourrions également écrire le terme résultant $\Lambda\alpha.(\lambda x^{\alpha}.\ (P^{\alpha\to W}x))\ (f^{U\to\alpha}\tau^{U})\{V\}$, mais l'abstraction au second ordre serait redondante.

2. Utilisation globale d'une transformation :

$$(\Lambda\alpha\lambda f^{\alpha\to V}.(\lambda x^{\alpha}.\ (P^{V\to W}(f^{\alpha\to V}x^{\alpha}))))\{U\}\ f^{U\to V}\ \tau^{U}$$

Supposant toujours que f est disponible chez P ou τ ; cette variation de l'application infère le type $\{U\}$ et le morphisme associé f de la formule originelle $(\lambda x^{V}.\ (P^{V\to W}x))\tau^{U}$. Cette construction donne des « places » pour l'application locale des transformations.

Ici, le type de l'argument x n'est pas connu, et le terme utilisé pour la représentation du prédicat doit le capturer avant d'inférer le domaine de la transformation f ; l'abstraction au second ordre n'est donc pas optionnelle dans cette construction.

5.3 Flexibilité

Comme on le verra plus tard, chaque transformation f est donnée dans le lexique avec *degré de flexibilité* $\delta(f)$. Chaque utilisation d'une transformation modifie également la flexibilité du nœud de l'analyse syntaxique dans laquelle elle se trouve, ce pour éviter les co-prédications hasardeuses. L'application a lieu dans les cas suivants :

- L'argument de la transformation est de degré 0 ou 1, et la transformation est de degré 1 : application locale.
- L'argument de la transformation est de degré 0 ou 1, et la transformation est de degré 2 : application globale.

5.4 Exemples

- Quand l'utilisation d'une transformation n'est pas nécessaire, l'application procède comme d'habitude – on peut dire également qu'il s'agit d'une transformation *Identité* du type concerné, de degré 1 (flexible). Ici, le type φ est celui des objets physiques :

petit caillou

$$\overbrace{\overbrace{(\lambda x^{\varphi}.\ (\text{petit}^{\varphi\to\varphi}\ x))}^{\text{small}}\ \tau^{\varphi}}^{\text{caillou}}$$
$$(\text{petit}\ \tau)^{\varphi}$$

- Dans l'exploitation de qualia, une information telle que la présence d'un quale *agentif* de type *personne*, associé à l'entrée lexicale de *sourire*, serait modélisée par la présence d'une transformation $f_a^{S\to P}$ (le terme est de degré 1 et peut être utilisé localement sans problème si rien ne s'y oppose dans le reste de la prédication ; il est associé au lexème *sourire* et son type *S*, et dénote la relation entre un sourire et son exécutant). Alors, cette transformation sera utilisée pour modéliser la coercition de type appropriée :

un sourire amical

$$\overbrace{\overbrace{(\lambda x^{P}.\ (\text{amical}^{P\to t}\ x))}^{\text{amical}}\ \tau^{S}}^{\text{sourire}}$$
$$(\text{amical}^{P\to t}\ x)))(f_a^{S\to P}\tau^{S})$$
$$(\text{amical}\ (f_a\ \tau))$$

- **Co-prédication incorrecte** : les opérations lexicales destructrices sont de degré 2, et ne peuvent être utilisées que globalement. Après une utilisation de cette transformation, et dans la même phrase, il est donc impossible de faire appel au type précédemment utilisé (transformation « Identité », de degré 1). Si on utilise une opération de *grinding* telle que $f_g^{Poisson\to Nourriture}$, il est donc impossible de trouver une dérivation correcte pour la phrase inélégante ci-dessous :

(? ?) *Le thon d'hier était bon nageur et délicieux.*

- **Co-prédication correcte** : la relation entre un mot et ses divers aspects compatibles entre eux peut être modélisée en utilisant des termes de degré 1, dont les transformations locales seront co-applicables. Soient $f_p^{V\to P}$ et $f_l^{V\to L}$ représentant les relations entre les mots dénotant une *ville* aux *personnes* et *lieux* associés, respectivement, toutes deux flexibles. Alors la phrase co-prédicative suivante est valide :

Copenhague est à la fois un port de mer et une capitale cosmopolite.

Intuitivement, il devrait y avoir une conjonction entre les prédicats $cospl^{P\to t}$, $\text{cap}^{V\to t}$ et $port^{L\to t}$, appliqués au même k^V (avec les types V pour ville, P pour personnes, L pour lieu, et la constante k Copenhague). Si $V = P = L = e$, comme se serait le cas dans la sémantique Montagovienne classique, alors on obtiendrait $(\lambda x^e(\text{and}^{t\to(t\to t)}((\text{and}^{t\to(t\to t)}\ (\text{cospl}\ x)\ (\text{cap}\ x))\ (\text{port}\ x)))\ k$. Ici, la solution canonique pour construire un terme bien typé à partir du *et*, des λ-termes principaux et optionnels est la suivante :

Si *et* est la constante habituelle de type $t \to (t \to t)$, le *et* entre deux prédicats $P^{\alpha\to t}$ et $Q^{\beta\to t}$ avec deux domaines différents α et β est :

$$\Lambda\alpha\Lambda\beta\lambda P^{\alpha\to t}\lambda Q^{\beta\to t}\Lambda\xi\lambda x^{\xi}\lambda f^{\xi\to\alpha}\lambda g^{\xi\to\beta}.(\text{et}\ (P\ (f\ x))(Q\ (g\ x)))$$

Ce terme lie tout d'abord des variables du second ordre aux types des domaines de P et Q, qui sont des α et β arbitraires ($\Lambda\alpha\Lambda\beta$). Le type de l'argument x, également arbitraire, est ensuite lié à ξ ($\Lambda\xi$). Pour appliquer cet argument aux deux prédicats, il faut donner deux emplacements séparés pour des transformations indépendantes de chaque argument (ce qui fonctionne quand elles sont de degré 1), suivant le type du domaine de chaque prédicat, et la conjonction est donc analogue à $(\text{et}\ (P\ (f\ x))\ (Q\ (g\ x)))$ (deux transformations de degré 1, locales). Nous avons alors directement le typage nécessaire pour chaque transformation ($\lambda f^{\xi\to\alpha}\lambda g^{\xi\to\beta}$), et donc le terme ci-dessus. Notons que, pour avoir un terme bien formé et bien typé, il faut que deux transformations de degré 1 soient disponibles pour se substituer à f et g.

La conjonction est d'abord appliquée aux types P et V et à $cospl^{P\to t}$ et $cap^{V\to t}$, donnant :

$$(*) \qquad \Lambda\xi\lambda x^{\xi}\lambda f^{\xi\to P}\lambda g^{\xi\to V}(\text{et}\ (\text{cospl}^{P\to t}\ (f\ x))\ (\text{cap}^{V\to t}\ (g\ x)))$$

De même, l'application de la même conjonction à $(*)$ et à $(\text{port}\ x)$ donne :

$$\Lambda\xi\lambda x^{\xi}\lambda f^{\xi\to P}\lambda g^{\xi\to V}\lambda h^{\xi\to L}(\text{et}\ (\text{et}\ (\text{cospl}^{P\to t}\ (f\ x))\ (\text{cap}^{V\to t}\ (g\ x)))\ (\text{port}^{L\to t}\ (h\ x)))$$

Ce terme est ensuite appliqué à type de l'argument, V, à l'argument, k^V, et aux morphismes. $cap^{V\to t}$ est déjà du type attendu, nous utilisons donc $id^{V\to V}$. Le résultat est un terme de type t :

$$(\mathrm{and}^{t\to(t\to t)}\ (\mathrm{and}^{t\to(t\to t)}\ (\mathrm{cospl}\ (f_p\ k^T)^P)^t)\ (\mathrm{cap}\ (\mathrm{id}\ k^T)^T)^t)^t\ (\mathrm{port}\ (f_l\ k^T)^L)^t)^t$$

Chapitre 6

Un processus intégré d'analyse des sens

Modifications nécessaires des processus analytiques

Pour intégrer ces différents mécanismes dans le tout cohérent formé par l'analyse montagovienne, il nous faut reprendre certaines parties de cette dernière. Notamment, il nous faut préciser la description du lexique et de la flexibilité des opérations lexicales, et définir complètement le processus de la composition. Alors seulement nous pourrons examiner la pertinence du système mis en place.

6.1 Description lexicale

Il s'agit tout d'abord de bien définir l'architecure d'un lexique tel qu'on le souhaite voir opérer. Chaque entrée est ainsi associée à un λ-terme principal de ΛTY_n et à une suite de couples décrivant une opération lexicale et le degré de flexibilité de cette dernière.

6.1.1 Le degré de flexibilité

Une opération lexicale f peut être :

interdite : elle ne peut en aucun cas avoir lieu, il s'agit d'une spécification technique. On note $\delta(f) = 0$.

flexible : elle peut avoir lieu sans restriction aucune. On note $\delta(f) = 1$.

semi-flexible : elle impose une restriction sur toutes les instances d'un même mot, mais pas sur les reprises anaphoriques, pronominales, discursives… associées. On note $\delta(f) = 2$.

rigide : elle impose une restriction absolue sur toutes les instances du concept associé. On note $\delta(f) = 3$.

6.1.2 Une entrée lexicale

Une entrée lexicale est la donnée du λ-terme principal, représentant la structure argumentale du lexème en termes de variables liées et typées dans ΛTY_n, et de l'ensemble des opérations que ce lexème peut contribuer de façon connue, chacune associée à son degré δ de flexibilité. Par exemple, *Paris* a pour entrée :

$$\left\langle \lambda x^{Ville}.(\text{Paris}^{Ville \to t}\ x)\ ;\ \underset{1}{Id = \lambda x^{Ville}.x},\ \underset{1}{f_P^{Ville \to Personnes}},\ \underset{1}{f_L^{Ville \to Lieu}},\ \underset{3}{f_G^{Ville \to Gouvernement}} \right\rangle$$

6.1.3 Le lexique

Le lexique est organisé de manière classique : à chaque mot correspond un ou plusieurs lexèmes, chacun associé à son entrée lexicale. Les mots à plusieurs entrées correspondent, non pas aux polysèmes, mais aux polymorphes accidentels (comme *bar*) ; ces différentes entrées sont différenciées par leur type élémentaire.

6.1.4 Mécanisme de récupération d'une entrée depuis la structure tectogrammatique

La problématique est la suivante : étant donnée une phrase analysée en syntaxe et sa structure tectogrammatique élémentaire, comment amorce-t-on la composition ? Le mécanisme supposé d'analyse doit récupérer, du lexique, les entrées correspondant à chaque mot. S'il y a une préférence manifeste de types pour les polymorphes, l'appliquer ; sinon, effectuer une recherche rapide sur les types-cible des opérations lexicales et les appareiller du mieux possible. Par défaut, engendrer deux calculs ou plus, sachant que (sauf double-sens) tous sauf un devraient échouer.

6.1.5 Utilisation des composantes de l'entrée dans la composition

On appareille ensuite chaque paire prédicat-argument selon le typage du prédicat. Dans la plupart des cas, cela doit résulter en une application simple et directe. Dans le reste des cas, il s'agit de trouver une opération lexicale permettant l'application, en cherchant dans chacune des composantes du prédicat et de l'argument si l'opération est applicable. On compare ensuite le niveau actuel de flexibilité de la cible (c'est-à-dire le degré de flexibilité des opérations déjà effectuées) avant d'appliquer ou non l'opération voulue (une opération de degré > 1 ne peut s'appliquer sur une cible de degré > 0). Puis on passe au nœud parent de l'arbre, en faisant remonter le degré de flexibilité suivant la tectogrammaire (et le ramenant à 0 en cas de degré < 3 au passage d'un nœud *S*), et en faisant descendre les éventuelles opérations sur des sous-arbres.

6.2 Quantification et individuation

Une application très particulière de la sémantique lexicale constiste en l'utilisation de quantificateurs successifs sur différents aspects d'un mot logiquement polysémique. Il s'agit de prendre en compte ces phénomènes très particuliers.

6.2.1 Les données

Les phénomènes à traiter sont, par exemple :

1. L'étudiant a lu tous les livres de la bibliothèque.
2. Tous les livres de la bibliothèque ont disparu dans son incendie.
3. Tous les livres de la bibliothèque ont disparu dans son incendie, mais l'étudiant les avait tous mémorisés.
4. L'étudiant a lu tous les livres de la bibliothèque, afin de vérifier qu'ils étaient en bon état.

L'intuition est que les ensembles en 1 et 2 sont différents, et qu'un mécanisme complexe est mis en œuvre pour les faire cohabiter au sein de 3. [Asher, 2011], qui traite régulièrement ces exemples, affirme que les conditions d'individuation et de comptage d'un type multifacettes sont spécifiques et nécessitent un mécanisme qui leur est propre ; [Jayez, 2008] argue que ces constructions ne sont pas unique dans le langage et que des circonstancielles telles que celle présente en 4 suffisent à effectuer ce genre de modification.

6.2.2 Proposition

Considérant que des phénomènes tels que la co-prédication peuvent être traités, dans notre formalisation, à l'aide d'opérateurs de conjonction bien définis (notamment un « et » du second ordre, permettant l'adaptation de chacun de ses membres aux types ciblés), il nous paraît naturel de donner un traitement basé sur des quantificateurs rendant compte des phénomènes observés plutôt que de chercher un mécanisme spécifique à la cible. Il apparaît en effet qu'il y a, dans les exemples cités, deux actions effectuées par le quantificateur :

Individuation

Un type est utilisé pour sélectionner l'ensemble des objets, de ce type, concerné par la quantification.

Projection

Un type est ensuite utilisé pour projeter (l'opération mathématique est surjective) l'ensemble ainsi construit en un ensemble d'individus utilisables. Si les deux types sont identiques, il s'agit de l'identité. Selon les prédications utilisées, il peut y avoir plusieurs types-cible.

La sémantique est donc :
Soient α le type de l'individuation, $\beta_1, \beta_2, \ldots$ les types-cible, ξ le type de l'argument. On aura, parallèlement, des prédicats P de sélection, $Q_1, Q_2, \ldots$ de prédication et A d'argument. Soit un quantificateur et un opérateur correspondant à la réalisation « classique » de cette quantification. On aura f le morphisme entre types correspondant à l'individuation et $g_1, g_2 \ldots$ ceux correspondant aux projections.

Alors (pour deux prédicats) le quantificateur universel avec projections s'écrit, avec ses types :

$$\Lambda\alpha\beta_1\beta_2\xi\,\lambda P^{\alpha\to t} Q_1^{\beta_1\to t} Q_2^{\beta_2\to t} A^{\xi\to t} f^{\xi\to\alpha} g_1^{\xi\to\beta_1} g_2^{\xi\to\beta_2} x^{\xi} \;/$$

$$\forall\, x\,.\,(\Rightarrow^{t\to t\to t}\ (\wedge^{t\to t\to t}\ (A^{\xi\to t}\ x^{\xi})^{t}\ (P^{\alpha\to t}\ (f^{\xi\to\alpha}\ x)^{\alpha})^{t})^{t}\ /$$

$$(\wedge^{t\to t\to t}\ (Q_1^{\beta_1\to t}\ (g_1^{\xi\to\beta_1}\ x^{\xi})^{\beta_1}\ (Q_1^{\beta_2\to t}\ (g_2^{\xi\to\beta_2}\ x^{\xi})^{\beta_2})^{t})^{t}$$

Emploi du quantificateur avec projections

Ce quantificateur porte sur trois aspects : un prédicat donnant les propriétés essentielles de l'argument, $A^{\xi\to t}$, un prédicat opérant une première sélection avec adaptation pour le comptage des individus, $P^{\alpha\to t}$, et les prédicats « cibles », $Q_1^{\beta_1\to t}, Q_2^{\beta_2\to t}\ldots$, qui portent la prédication proprement dite et permettent chacun une adaptation. Ce mécanisme permet de sélectionner les individus sur lesquels va s'opérer la quantification selon un type qui peut être différent de ceux utilisés dans la prédication finale.

Autre possibilité

Suivant l'ordre des prédicats utilisés, il est possible de donner une autre sémantique aux quantifications avec projections : utiliser tout d'abord la quantification « classique », avec une prédication comportant une adaptation rigide, puis utiliser, lors de la seconde prédication, un terme permettant de remonter dans les adaptations déjà utilisées afin de reprendre une prédication correcte. Ainsi, si on effectue d'abord la prédication $(P\ (f\ x))$ et qu'on veut effectuer ensuite $(Q\ (g\ x))$, on utilise le terme $\lambda wPQ\,.\,(\wedge\ (Q\ (g\ (f^{-1}\ w)))\ (P\ w))$.

La difficulté de cette approche est d'effectuer correctement les reprises anaphoriques pour remonter au bon terme ; il serait nécessaire, pour la mettre en œuvre, de maintenir un contaxte dynamique comportant les prédications déjà utilisées et les termes accessibles. Son intérêt est de donner une modélisation différente du processus d'adaptation, dans laquelle chaque nouvelle information vient modifier un état informationnel déjà établi ; la pertinence de ces deux modélisations pourrait être évaluée par des tests cognitifs.

Illustrations

Dans les exemples précédents, nous avons (avec *Livre* comme type de départ, disposant de deux morphismes respectivement vers I, type des contenus informatifs, et φ, type des objets physiques) :

1. Type d'individuation : I. Type de projection : I.
2. Type d'individuation : φ. Type de projection : φ.
3. Type d'individuation : φ. Type de projection : I.
4. Type d'individuation : I. Type de projection : φ.

6.3 Architecture d'un analyseur

L'analyse telle que nous l'envisageons se déroule donc suivant les étapes suivantes :

1. Analyse graphique ou phonologique pour produire le texte de départ.

2. Analyse syntaxique, donnant la structure tectogrammatique du texte.

3. Composition, étape par étape, suivant la structure tectogrammatique. Chaque étape inclut :

   - L'application sur un nœud le plus bas possible dans la structure du prédicat à l'argument, considéré de degré 1 par défaut.
   - Si le type des lexèmes le permet, une réduction directe.
   - Si le type des lexèmes permet une application avec transformation, la déduire et réduire, en changeant si nécessaire le degré du nœud résultant.
   - Sinon, échouer.
   - Reprendre jusqu'à réduire entièrement le terme.

4. Restitution de la forme logique du texte.

5. Interprétation et utilisation du résultat.

6.4 Analyses multiples

Entre la sélection du « bon » lexème dans le lexique et celle de la transformation « correcte » dans l'application, il y a plusieurs possibilités pour que les analyses menant à une forme logique bien typée ne soit pas uniques. Ceci, nous le verrons, est plus un bénéfice qu'un problème ; toujours est-il que notre système doit pouvoir gérer les analyses multiples d'un même texte (problème qui se prête assez naturellement à la parallélisation).

Chapitre 7

Complétude et apports de l'approche

7.1 La réponse aux phénomènes posés

Notre approche, bien plus légère que [Pustejovsky, 1995] dans son ensemble, couvre la majorité des phénomènes de polysémie logique : revenons sur les exemples précédemment évoqués…

7.1.1 Prédication normale

Le système se comporte comme l'analyse traditionnelle dans le cas d'une prédication « normale ».

Un grand rocher

La structure profonde de cette prédication peut être interprétée comme

$$(\text{grand rocher})$$

Nous disposons d'une entrée lexicale pour *grand* :

$$\left\langle \Lambda\xi\,\lambda f^{\xi\to\varphi}\lambda x^{\varphi}.(\text{grand}^{\varphi\to t}\ (f\ x))\ ;\ \begin{matrix} Id=\lambda x^{\varphi}.x \\ 1 \end{matrix} \right\rangle$$

Et d'une pour *rocher* :

$$\left\langle \text{rocher}^{\varphi}\ ;\ \begin{matrix} Id=\lambda x^{\varphi}.x \\ 1 \end{matrix} \right\rangle$$

La composition des structures argumentales (les λ-termes principaux de chaque entrée) se fait sans problème de typage, et donc sans utilisation de transformation (permise par la présence de transformations identité) :

$$\Lambda\xi\,\lambda f^{\xi\to\varphi}\lambda x^{\xi}.(\text{grand}^{\varphi\to t}\ (f\ x))\ \{\varphi\}\mathbf{Id}^{\varphi\to\varphi}\ \text{rocher}^{\varphi}$$

$$(\text{grand}^{\varphi\to t}\ (Id\ \text{rocher}))$$

$$(\text{grand rocher})^{t}$$

7.1.2 Exploitation de qualia

En cas *d'exploitation de qualia*, une transformation de l'argument est utilisée directement.

Une épée efficace

Soit une structure profonde, quantification mise à part, de :

$$(\text{efficace epee})$$

Le prédicat *efficace* porte sur les événements :

$$\left\langle \Lambda\xi\,\lambda f^{\xi\to Evt}\lambda x^{\xi}.(\text{efficace}^{(Evt\to t}\ (f\ x))\ ;\ \underset{1}{Id = \lambda x^{Evt}.x} \right\rangle$$

Le lexème *épée* est donné avec ses quatre *qualia*, sous forme de transformations.

$$\left\langle \text{epee}^{Arme}\ ;\ \underset{1}{Id = \lambda x^{Arme}.x}\ \underset{1}{f^{Arme\to Artefact}_{Formel}}\ \underset{1}{f^{Arme\to Metal}_{Const}}\ \underset{1}{f^{Arme\to Evt}_{Telique}}\ \underset{1}{f^{Arme\to A}_{Agent}} \right\rangle$$

Pour être élégante, la composition doit faire appel à une transformation vers le type des événements ; le degré de la transformation (1) permet cette utilisation :

$$\Lambda\xi\,\lambda f^{\xi\to Evt}\lambda x^{\xi}.(\text{efficace}^{(Evt\to t}\ (f\ x))\ \{Arme\}\ \mathbf{f}^{\mathbf{Arme\to Evt}}_{\mathbf{Telique}}\ \text{epee}^{Arme}$$

$$(\text{efficace}\ (f_{Telique}\ \text{epee}))$$

7.1.3 Résultatifs

Le mécanisme est le même que pour celui de l'exploitation de qualia.

La construction est au bout de la rue.

En simplifiant, la structure est la suivante :

$$(\text{situation construction})$$

Le prédicat *au bout de la rue* (situation) est une description appliquée à un objet physique :

$$\left\langle \Lambda\xi\,\lambda f^{\xi\to\varphi}\lambda x^{\xi}.(\text{situation}^{(\varphi\to t}\ (f\ x))\ ;\ \underset{1}{Id = \lambda x^{\varphi}.x} \right\rangle$$

L'argument est un résultatif, comprenant donc une transformation en résultat. Nous devons également inclure une relation hyponimique vers les objets physiques pour une dérivation complète ; cette dernière transformation est induite par la hiérarchie ontologique des types, et reste présente dans le lexique (comme celle liant le type processus au type des événements).

$$\left\langle \text{construction}^{Processus}\ ;\ \underset{1}{Id = \lambda x^{Processus}.x}\ \underset{1}{f_{res}^{Processus\to Resultat}}\ \underset{1}{f_{Hevt}^{Processus\to Evt}}\ \underset{1}{f_{Hphy}^{Resultat\to\varphi}} \right\rangle$$

La transformation nécessaire lors de l'application est la composée de celle donnant le résultat et de celle induite par la relation hyponimique correspondante :

$$\Lambda\xi\,\lambda f^{\xi\to\varphi}\lambda x^{\xi}.(\text{situation}^{(\varphi\to t}\ (f\ x))\ \{Processus\}\ \mathbf{f_{Hphy}}\cdot\mathbf{f_{res}}\ \text{construction}^{Processus}$$

$$(\text{situation}\ (f_{Hphy}\ (f_{res}\ \text{construction})))$$

7.1.4 Objets multifacettes

Là encore, le mécanisme est identique.

Le livre est lourd.
Le livre est intéressant.

Un *livre* est un objet multifacettes complexe, de type *Lisible* (comme les articles, journaux, revues, preuves), comportant deux transformations vers ses facettes physiques et informationnelles.

$$\left\langle \text{livre}^{Lisible} \; ; \; \begin{matrix} Id = \lambda x^{Lisible}.x & f_{phy}^{Lisible\to\varphi} & f_{info}^{Lisible\to Info} \\ 1 & 1 & 1 \end{matrix} \right\rangle$$

Les prédicats utilisés sont ordinaires :

$$\left\langle \Lambda\xi\lambda f^{\xi\to\varphi}\lambda x^{\xi}.(\text{lourd}^{\varphi\to t}\ (f\ x)) \; ; \; \begin{matrix} Id = \lambda x^{\varphi}.x \\ 1 \end{matrix} \right\rangle$$

$$\left\langle \Lambda\xi\lambda f^{\xi\to Info}\lambda x^{\xi}.(\text{interessant}^{Info\to t}\ (f\ x)) \; ; \; \begin{matrix} Id = \lambda x^{Info}.x \\ 1 \end{matrix} \right\rangle$$

Selon la prédication utilisée, il suffit de sélectionner la transformation appropriée donnée dans *livre*.

$$\Lambda\xi\lambda f^{\xi\to\varphi}\lambda x^{\xi}.(\text{lourd}^{(\varphi\to t}\ (f\ x))\{Lisible\}\ \mathbf{f_{phy}}\ \text{livre}^{Lisible}$$

$$(\text{lourd}\ (f_{phy}\ \text{livre}))$$

Et :

$$\Lambda\xi\lambda f^{\xi\to Info}\lambda x^{\xi}.(\text{interessant}^{(Info\to t}\ (f\ x))\{Lisible\}\ \mathbf{f_{info}}\ \text{livre}^{Lisible}$$

$$(\text{interessant}\ (f_{info}\ \text{livre}))$$

7.1.5 Co-prédications inélégantes

C'est dans le cas des co-prédications que le degré de flexibilité des transformations devient utile, notamment pour déterminer l'élégance de ces phrases complexes. Pour les co-prédications inélégantes, une des transformations utilisées est trop *rigide* pour collaborer avec d'autres (y compris l'identité), et la dérivation échoue. Deux exemples déjà longuement évoqués :

? *Ce saumon était très rapide et délicieux.*

? *Paris est une ville aux bâtiments élégants et a refusé d'attaquer l'Irak.*

Dans le premier cas, nous avons affaire à deux prédicats portant respectivement sur un animal (le reste de l'entrée de *rapide* étant omis par soucis de simplicité) et de la nourriture.

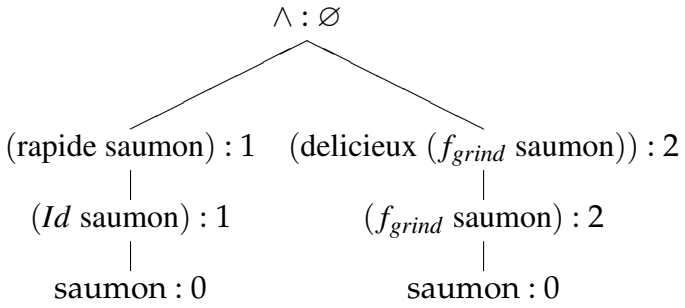
$$\left\langle \Lambda\xi\,\lambda f^{\xi\to Animal}\lambda x^{\xi}.(\text{rapide}^{(Animal\to t}\ (f\ x))\ ;\ \begin{matrix} Id=\lambda x^{Animal}.x \\ 1 \end{matrix} \right\rangle$$

$$\left\langle \Lambda\xi\,\lambda f^{\xi\to Nourriture}\lambda x^{\xi}.(\text{delicieux}^{(Nourriture\to t}\ (f\ x))\ ;\ \begin{matrix} Id=\lambda x^{Nourriture}.x \\ 1 \end{matrix} \right\rangle$$

L'argument est un animal susceptible de subir une opération de *grinding* (par préparation culinaire), et dispose donc d'une transformation semi-rigide correspondante.

$$\left\langle \text{saumon}^{Animal}\ ;\ \begin{matrix} Id=\lambda x^{Animal}.x & f_{grind}^{Animal\to Nourriture} \\ 1 & 2 \end{matrix} \right\rangle$$

La prédication consisterait à appliquer successivement la transformation semi-rigide et la transformation identité, ce qui donnerait lieu à un conflit dans les degrés de flexibilité :



(Cette dérivation est symétrique).

Dans le second cas, la situation est identique. On dispose de deux prédicats, l'un portant sur les villes, l'autre sur les gouvernements :

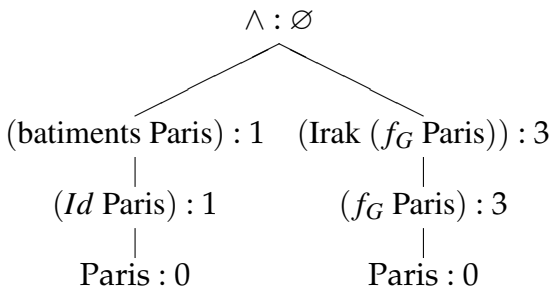
$$\left\langle \Lambda\xi\,\lambda f^{\xi\to Ville}\lambda x^{\xi}.(\text{batiments}^{(Ville\to t}\ (f\ x))\ ;\ \begin{matrix} Id=\lambda x^{Ville}.x \\ 1 \end{matrix} \right\rangle$$

$$\left\langle \Lambda\xi\,\lambda f^{\xi\to Gouvernement}\lambda x^{\xi}.(\text{Irak}^{(Gouvernement\to t}\ (f\ x))\ ;\ \begin{matrix} Id=\lambda x^{Gouvernement}.x \\ 1 \end{matrix} \right\rangle$$

L'entrée lexicale de *Paris* a déjà été donnée :

$$\left\langle \lambda x^{Ville}.(\text{Paris}^{Ville\to t}\ x)\ ;\ \underset{1}{Id = \lambda x^{Ville}.x}\ ,\ \underset{1}{f_P^{Ville\to Personnes}}\ ,\ \underset{1}{f_L^{Ville\to Lieu}}\ ,\ \underset{3}{f_G^{Ville\to Gouvernement}} \right\rangle$$

Et, comme dans le cas précédent, il est impossible d'appliquer l'identité et une transfmormation rigide simultanément :



(Là aussi, la dérivation ne dépend pas de l'ordre de résolution.)

7.1.6 Co-prédications élégantes

Il y a de très nombreuses manières de produire des co-prédications élégantes. Une utilisation simultanée de multiples transformations flexibles fonctionne, tout comme l'utilisation d'une transformation semi-flexible dans des conditions discursives adaptées ; nous examinerons ces deux cas.

un livre lourd mais intéressant.
Ce saumon est délicieux. Il était très rapide.

Dans le premier cas, il s'agit de reprendre les deux prédicats déjà donnés précédemment :

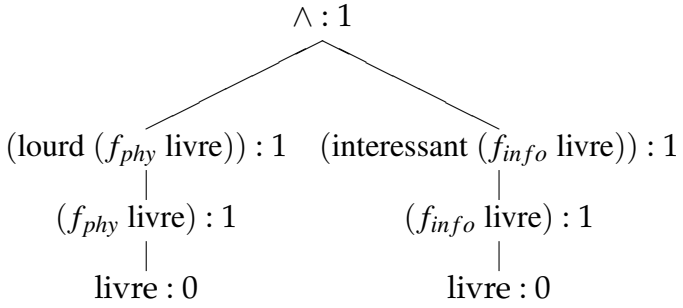
$$\left\langle \Lambda\xi\,\lambda f^{\xi\to\varphi}\lambda x^{\xi}.(\text{lourd}^{\varphi\to t}\ (f\ x))\ ;\ \underset{1}{Id = \lambda x^{\varphi}.x} \right\rangle$$

$$\left\langle \Lambda\xi\,\lambda f^{\xi\to Info}\lambda x^{\xi}.(\text{interessant}^{Info\to t}\ (f\ x))\ ;\ \underset{1}{Id = \lambda x^{Info}.x} \right\rangle$$

L'argument, déjà donné également, est le suivant :

$$\left\langle \text{livre}^{Lisible}\ ;\ \underset{1}{Id = \lambda x^{Lisible}.x}\ \underset{1}{f_{phy}^{Lisible\to\varphi}}\ \underset{1}{f_{info}^{Lisible\to Info}} \right\rangle$$

Il s'agit de vérifier que les transformations peuvent s'appliquer simultanément :



La dérivation est donc possible. Formellement, il s'agit de :

$$\Lambda\xi\,\alpha\beta\lambda f^{\xi\to\alpha} g^{\xi\to\beta}\; P^{\alpha\to t} Q^{\beta\to t}\;.\;(\wedge\;(P\;(f\;x))\;(Q\;(g\;x)))\;/$$
$$\{Lisible\}\;\{\varphi\}\;\{Info\}\;\mathbf{f_{phy}}\;\mathbf{f_{info}}\;\text{lourd}^{\varphi\to t}\;\text{interessant}^{Info\to t}\;\text{livre}^{Lisible}$$
$$(\wedge\;(\text{lourd}\;(f_{phy}\;\text{livre}))\;(\text{interessant}\;(f_{info}\;\text{livre})))$$

Pour le deuxième cas, il s'agit de reprendre les mêmes prédications :

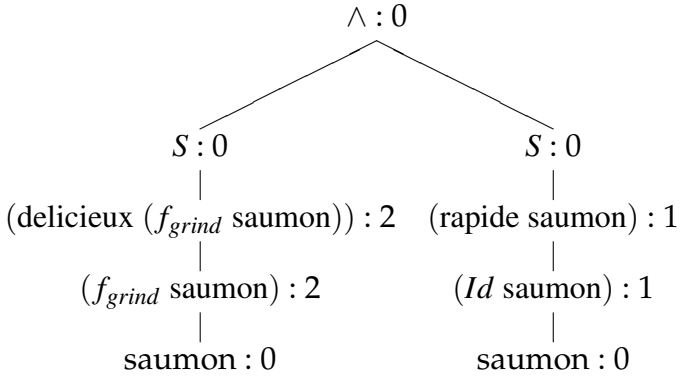
$$\left\langle \Lambda\xi\lambda f^{\xi\to Animal}\lambda x^{\xi}.(\text{rapide}^{(Animal\to t}\;(f\;x))\;;\;\underset{1}{Id = \lambda x^{Animal}.x}\right\rangle$$

$$\left\langle \Lambda\xi\lambda f^{\xi\to Nourriture}\lambda x^{\xi}.(\text{delicieux}^{(Nourriture\to t}\;(f\;x))\;;\;\underset{1}{Id = \lambda x^{Nourriture}.x}\right\rangle$$

Et le même argument, avec sa transformation semi-rigide :

$$\left\langle \text{saumon}^{Animal}\;;\;\underset{1}{Id = \lambda x^{Animal}.x}\;\;\underset{2}{f_{grind}^{Animal\to Nourriture}}\right\rangle$$

Cette fois, la flexibilité est garantie par la reprise anaphorique – ici, le passage d'un nœud S :



La dérivation est donc possible, avec la même structure que précédemment :

$$\Lambda\xi\,\alpha\beta\lambda\, f^{\xi\to\alpha} g^{\xi\to\beta}\ P^{\alpha\to t} Q^{\beta\to t}\ .\ (\wedge\ (P\ (f\ x))\ (Q\ (g\ x)))\ /$$
$$\{Animal\}\ \{Nourriture\}\ \{Animal\}\ \mathbf{f_{grind}}\ \mathbf{Id}\ \text{delicieux}^{Nourriture\to t}\ \text{rapide}^{Animal\to t}\ \text{saumon}^{Animal}$$
$$(\wedge\ (\text{delicieux}\ (f_{grind}\ \text{saumon}))\ (\text{rapide saumon}))$$

7.1.7 Quantifications co-prédicatives

Le but est ici, non seulement de vérifier l'élégance des phrases concernées, mais également de prédire correctement leur sens dans les domaines de portée des quantificateurs.

Tous les livres de la bibliothèque ont brûlé, mais je les avais déjà lus.

La forme profonde de la phrase peut être approximée comme suit (notre quantificateur prend un nombre arbitraire de prédicats en argument) :

$$\forall x\ .\ (livre\ x)\ (bibliotheque\ x)\ (brule\ x)\ (lu\ x)$$

En utilisant le quantificateur avec projections vu auparavant, nous avons donc, comme terme de départ :

$$\Lambda\alpha\beta_1\beta_2\xi\lambda A^{\xi\to t} P^{\alpha\to t} Q_1^{\beta_1\to t} Q_2^{\beta_2\to t} f^{\xi\to\alpha} g_1^{\xi\to\beta_1} g_2^{\xi\to\beta_2} x^{\xi}\ ./$$
$$(\text{set}_\forall x\ (\Rightarrow\ (\wedge\ (A\ x)\ (P\ (f\ x)))\ (\wedge\ (Q_1\ (g_1\ x))\ (Q_2\ (g_2\ x))))))\ /$$
$$\{\varphi\}\ \{I\}\ \{\varphi\}\ \{Lisible\}\ (\lambda x^{Lisible}.(\text{livre}\ x))\ (\lambda x^{\varphi}.(\text{bibliotheque}\ x))\ /$$
$$(\lambda x^{I}.(\text{lu}\ x))\ (\lambda x^{\varphi}.(\text{brule}\ x))\ f_{phy}^{Lisible\to\varphi}\ f_{info}^{Lisible\to I}\ f_{phy}^{Lisible\to\varphi}\ x^{Lisible}$$

En effet :

- La quantification porte sur des livres, c'est-à-dire des objets de type *Lisible* qui vérifient le prédicat $\lambda x.(\text{livre}\ x)$.
- Elle sélectionne les objets quantifiés selon une propriété des objets de type φ, $\lambda x.(\text{bibliotheque}\ x)$.
- Elle a pour but de fournir des prédicats portants respectivement sur les objets de type I et de type φ, $\lambda x.(\text{lu}\ x)$ et $\lambda x.(\text{brule}\ x)$.

On peut vérifier la cohérence des types de ce terme, et le réduire en :

$$set_\forall x^{Lisible}\ (\Rightarrow\ (\wedge\ (\text{livre}\ x)\ (\text{bibliotheque}\ (f_{phy}\ x)))\ (\wedge\ (\text{lu}\ (f_{info}\ x))\ (\text{brule}\ (f_{phy}\ x))))$$

Il s'agit bien du résultat désiré.

7.2 Comparaison avec les autres approches

7.2.1 Pustejovsky

L'approche présentée ici ne contredit ni [Pustejovsky, 1991], ni [Pustejovsky, 1995]. Elle essaie de proposer un cadre logique dans lequel ces théories puissent être mises en place.

Cependant, [Pustejovsky, 1991] est plus fondé sur la nécesssité d'engrenger des travaux allant dans la direction d'une approche générative du lexique, et se contente de donner des exemples pratiques (très détaillés) de constructions ; il s'agit d'un argumentaire en faveur de développements futurs. [Pustejovsky, 1995] contiendra certains de ces développements, notamment une argumentation renouvellée et des discussions couvrant de multiples aspects de la linguistique, et l'introduction d'une réflexion sur les objets complexes.

Malgré toutes les qualités littéraires de la proposition originale de Pustejovsky, il faudra attendre des travaux subséquents pour combler la liaison avec les principes de compositionalité logique inhérents à l'analyse Montagovienne ; le présent manuscrit peut être considéré comme l'un d'entre eux.

7.2.2 Pinkal & Kohlhase

[Pinkal and Kohlhase, 2000] constitue une approche complète et novatrice du traitement des objets complexes. La différence principale avec la notre est le traitement utilisant une logique de traits avec ordre partiel : nous pensons que le calcul proposé est délicat à mettre en œuvre, car les opérations ne dépendent que des types concernés.

L'approche suivie est complète pour les objets de type complexe, et ne se distingue de la notre que par le traitement effectué et le choix de la logique.

Cependant, à notre connaissance, seuls l'article original et [Jacquey, 2001] ont traité de ce calcul.

Nous considérons être allés plus loin dans l'application et la recherche d'un formalisme, même si le notre manque de l'élégance de la publication de Pinkal & Kohlhase.

7.2.3 Asher

[Pustejovsky and Asher, 2000] était la première tentative de Asher et Pustejovsky pour donner une logique au lexique génératif, et principalement aux objets complexes. Ce premier effort fut poursuivi par [Asher and Pustejovsky, 2005] et [Asher, 2006], jusqu'à [Asher, 2008].

Cette démarche est très rapprochée de nos propres préoccupations : il s'agit de donner une logique et un calcul permettant aux notions introduites dans [Pustejovsky, 1995], dans le but d'intégrer à l'analyse sémantique Montagovienne les principes du lexique génératif. Cependant, les opérations les plus élémentaires de cette logique firent longtemps l'objet de critiques quant à la possibilité de leur mise en œuvre.

A Web of Words

Une réflexion constante autour de ces critiques, portant sur la refondation des principes de cette logique, a abouti a un travail de synthèse publié dans [Asher, 2011].

Il s'agit d'un travail de plusieurs années, fondé sur la collaboration de nombreux acteurs. Le raffinement du calcul au fur et à mesure des publications successives donne un système complexe et riche, étayé par de nombreux exemples de phénomènes très variés ; il s'agit d'une étape majeure dans la recherche sur la sémantique lexicale et ses modèles formels. En effet, si nous avons déjà évoqué les multiples travaux effectués ponctuellement par de nombreux auteurs, [Asher, 2011] propose le premier point synthétique et la première théorie complète dans le domaine depuis [Pustejovsky, 1995].

Bien que prenant appui sur l'ensemble des travaux précédents, cet ouvrage ouvre une nouvelle direction d'études en cherchant à ramener l'ensemble des phénomènes étudiés à un unique traitement logique, celui des présuppositions. Cherchant à faire apparaître cet élément de départ, l'auteur est amené à la description d'une théorie élaborée, se rapprochant des traitements du discours ou du dialogue. Il s'agit également d'une rupture avec les publications précédentes, ayant l'ambition d'apporter une réponse à chaque question soulevée par les nombreux critiques de l'approche de Pustejovsky, et de compléter les travaux ultérieurs, afin de permettre à ses successeurs de se concentrer sur l'implémentation de solutions concrètes. Si [Asher, 2011] ne parvient pas à satisfaire tous ses objectifs et nécessite de nombreuses clarifications (comme nous espérons que le présent manuscrit en apporte), les travaux de recherches qu'il représente sont considérables.

Comme [Pustejovsky, 1995], [Asher, 2011] est fondé sur une modélisation des concepts par une hiérarchie de types. Les contraintes sémantiques, qu'elles proviennent du lexique, de la syntaxe ou du discours, y sont représentées par des présuppositions de types, un ensemble d'opérations sur les types qui sont alors fonction de leur environnement. Une logique complète fondée sur la possibilité de contraindre, adapter, et composer ces types selon les besoins, et de créer des types complexes pour certains objets, est développée sur la base d'opérations sur les catégories. [Asher, 2011] soutient une approche philosophique forte, celle que la langue, en tant que représentation de la réalité, contient une information très riche, qu'on ne peut réduire à la simple référence ; chacune des constructions est justifiée et discutée, et prend appui sur de nombreux exemples.

Nous nous devons ici de présenter un bref résumé des idées principales de [Asher, 2011] : le principe et le mécanisme des présuppositions, la logique de compositions de types et la vision du monde soutenues dans l'ouvrage, avant de comparer notre approche à cette dernière, en exposant points forts, points de convergence et points faibles de cette théorie par rapports à celle défendue dans le présent manuscrit.

Résumé

Présuppositions
L'innovation principale de [Asher, 2011] est l'utilisation de présuppositions généralisées en tant que mécanisme prévisionnel pour la sémantique.

Phénomènes ciblés
Au départ de cette réflexion figure une interrogation sur de très nombreux cas de prédications.

Reprenant les interrogations de [Quine, 1960] et de très nombreux cas subséquents, l'ouvrage reprend et critique de multiples modèles qui l'ont précédé. Faisant le choix d'une sémantique lexicale riche, et rejetant l'approche de l'énumération des sens lexicaux par des arguments proches de ceux de [Pustejovsky, 1995] et plusieurs exemples concrets de polysémie relationnelle, l'auteur s'intéresse également aux modèles de Nunberg ou de Kleiber. Pour chacun, il détaille les manquements pilosophiques et logiques, et effectue également une critique mesurée de [Pustejovsky, 1995], et notamment de son système de calcul par trop simpliste.

Cette littérature abondante permet d'établir une phénoménologie très complète. Les études de cas sont nombreuses, et l'ouvrage s'appuie fortement sur les exemples utilisés par ses prédecesseurs ou des variantes.

De plus, les précédentes publications de l'auteur sur le sujet donnent un ensemble d'éléments supplémentaires.

Les phénomènes étudiés sont semblables à ceux que nous couvrons : il s'agt de cas de prédication avec adaptation, dans lesquels une information afférente à l'un des membres du couple prédicat-argument modifie la sémantique opérationnelle de l'autre membre de ce couple. Dans [Asher, 2011], la couverture est très large, de par l'accumulation de données sur le sujet : ainsi, on donne une typologie de verbes dénommés « coercitifs », tels *aimer*, *apprécier*, *commencer* (portant sur un aspect événementiel). Les phénomènes étudiés comportent également des applications d'adjectifs ou d'adverbes, et de multiples autres applications sémantiques.

L'auteur s'attache également à présenter des exemples issus de plusieurs langues, dont l'Anglais, le Français, le Chinois, le Japonais ou le Basque. En effet, l'étude a pour but une couverture universelle des phénomènes de prédication, et recherche donc à expliquer des phénomènes qui pourraient être occultés par les spécificités de l'une ou l'autre langue ; ces exemples permettent ainsi d'examiner les classificateurs ou d'autres marques d'appartenance à une catégorie sémantique particulière pour justifier les théories présentées.

Enfin, [Asher, 2011] ne se contente pas de présenter les prédications les plus simples, mais étudie de multiples constructions guidées par la syntaxe, dont le génitif. Le système de calcul peut ainsi s'appliquer à des phrases relativement élaborées.

Contextes

Le canevas présenté intègre aussi bien le contexte immédiat de chaque mot que les informations apportées par le reste de la phrase et du discours.

Le prédicat peut forcer à considérer l'argument sous un certain aspect : c'est l'adaptation du sens lexical, appelée ici « coercition ». De même, l'argument peut apporter des informations qui viendront modifier le sens du prédicat, et ce, dans de multiples constructions syntaxiques. L'étude se fonde donc, tout d'abord, sur l'évaluation d'apports du contexte immédiat à la sémantique d'un mot.

Dans une même phrase, un même discours, les modifications induites par la coercition peuvent se répercuter en de multiples instances, notamment en cas de référence anaphorique ; il est donc nécessaire de propager les modifications du contexte immédiat au contexte discursif. Un autre argument est que certaines constructions dans le discours, certains actes de langage, ou même de simples adjoints peuvent modifier la sémantique d'un mot ; ce dernier ne pouvant être conçu comme isolé, il reçoit de multiples contraintes sur son sens. L'ouvrage étudie donc également de multiples constructions de la phrase ou du discours qui, sans être des prédications directes, influent sur la sémantique.

De nombreux auteurs séparent l'analyse du discours de la sémantique, arguant que ces aspects relèvent de la pragmatique et ne doivent pas être traités de la même façon. [Asher, 2011], *a contrario*, affirme que certains aspects de la pragmatique, ainsi que les éléments issus du discours, doivent être intégrés au même niveau que l'analyse de la sémantique ; en effet, le sens de certaines prédications ne peut être connu sans informations pragmatiques supplémentaires. Les cas d'études comprennent donc également quelques exemples qui changeront de sémantique en fonction des paramètres d'énonciation.

L'ensemble de ces exemples est apporté pour montrer qu'un mécanisme unique est à l'œuvre dans tous les cas, et qu'il est, par bien des aspects, identique à celui de la coercition : la contrainte d'aspect. Il s'agit d'une information issue de la sémantique lexicale et de la syntaxe, qui force une donnée de la sémantique du mot étudié. Le problème est donc le même que dans les prédications avec adaptation : il s'agit d'étudier quel mécanisme peut être employé pour considérer l'objet étudié sous l'angle que lui impose le contexte.

Mécanisme général

L'ouvrage propose un mécanisme général pour ces apports d'information : la présupposition.

Dans ce mécanisme, la coercition est conçue comme une présupposition d'information : le prédicat coercitif « présuppose » un certain aspect pour son argument. Cette présupposition conduit à l'extraction d'informations sémantiques de l'argument portant sur l'aspect présupposé, et vient à échouer si cette information n'est pas disponible. C'est ainsi une forme de prédiction qui s'effectue lors de la prédication.

Les apports d'information provenant de la phrase, du discours ou d'un contexte pragmatique peuvent être considérés de la même manière : chacun apporte une ou plusieurs présuppositions sur la sémantique que réalise le mot, et ce dernier doit pouvoir se comporter d'une manière à satisfaire l'ensemble de ces présuppositions. Chaque apport d'information extérieure est donc vu comme une contrainte supplémentaire apportée à un mot, qui doit être compatible avec la sémantique lexicale de ce dernier.

Cette présupposition d'information sémantique, généralisée à l'ensemble des phénomènes étudiés, est ici représentée par un mécanisme de contraintes sur les types, qui sont donc les vecteurs des contraintes et des données apportées par les diverses sources.

Par conséquent, la sémantique de la phrase comporte un flux d'information transmis par les types en sus des prédications habituelles données par la structure de la syntaxe. Chaque lexème employé peut contribuer à cette information, qui modifie le sens des lexèmes qui suivront.

Logique de composition de types

Pour traiter l'ensemble des données apportées par les présuppositions, l'auteur développe un système logique complet appelé « logique de composition de types ».

Notion de types

Les types sont à la base du calcul proposé, guidant l'ensemble des prédications ; dans ce système, ils peuvent également faire l'objet de prédicats et de contraintes.

Chacun des termes dispose d'un type, ou ensemble de types, donné par le lexique. Ces types sont raffinés, basés sur une ontologie de concepts décrivant le monde et hiérarchisés : au niveau le plus bas, il existe un type correspondant à chaque mot. Cette ontologie permet d'associer à chaque lexème un type général qui peut être contraint, modifié par les informations provenant du contexte.

Les présuppositions d'information sémantique induites par les divers éléments du contextes se traduisent donc par des contraintes sur les types, représentées sous la forme de prédicats portants sur ces derniers. Les types sont alors considérés comme des variables et dépendent des affirmations dont ils sont l'objet.

Ainsi, on peut affirmer qu'un type descend d'un autre, est inclus dans un ensemble de types, ou encore est contenu dans un aspect donné de la sémantique lexicale d'un terme ; les prédicats de types comprennent également l'affirmation de certaines relations discursives comme la description ou l'élaboration. Les présuppositions peuvent apporter des prédicats très variés sur les types, qui s'intègrent au calcul de la forme logique de la phrase ; à leur tour, ces types contraignent les prédicats et arguments issus de la représentation directe des mots du texte.

Les multiples prédicats sur les types apportent ainsi autant de prévisions sur la signification finale de ces derniers. Par unification de variables sur ces prédicats, le calcul fixe ensuite chaque type, et la forme logique de la phrase est donnée par le système logique dans la tradition de la grammaire de Montague. Qu'ils soient complexes, dépendants, ou modifiés par une certaine construction précisée dans l'ouvrage, tous les apports d'informations sur les types font l'objet d'une opération de *justification* qui occasionne des modifications sur la structure de la forme logique de la phrase.

Types complexes

Certains lexèmes comportent une ambiguïté de types du point de vue de l'ontologie ; une grande partie de [Asher, 2011] est consacré au traitement de ces objets complexes, faisant appel à des mécanismes très particuliers.

Les objets « complexes » comprennent les mots comme *livre* (comprenant un aspect informationnel et un aspect physique) ou *déjeuner* (comportant un aspect physique et un aspect événementiel). Au premier abord, il est possible d'attribuer à ces lexèmes deux types différents, voire davantage, qui ne sont pas compatibles. La hiérarchie ontologique de types ne donne pas de solution immédiate pour traiter ces lexèmes d'un point de vue logique.

Adoptant un point de vue proche de [Pustejovsky, 1995], [Asher, 2011] considère que les objets complexes disposent simultanément de tous les types qu'ils peuvent assumer.

La difficulté de cette approche est de donner une représentation logique à ces types complexes. L'auteur examine de multiples propositions données dans la littérature, avant de les rejeter après analyse : il argue ainsi qu'un type complexe ne peut être issu de l'intersection des types qui les compose, non plus que de l'ensemble de ces derniers. En s'appuyant sur l'examen de phénomènes comme les co-prédications, l'auteur rejette également la solution de l'instanciation de types d'ordre supérieur.

La proposition de solution porte sur une construction spécifique, les types pointés, dont la notation est issue de [Pustejovsky, 1995]. Son principe est que les composants du type complexe sont liés à ce dernier par un ensemble de relations, dont l'implémentation est décrite par une logique catégorielle.

Logique catégorielle

La composition de types s'appuie sur une logique des catégories.

L'auteur examine les logiques précédemment étudiées et conclut à leur inadéquation, notamment pour la logique du premier ordre décrite dans [Pustejovsky, 1995], trop peu expressive.

Les types étant des objets élémentaires du système de calcul, ils disposent d'un format particulier. On distingue : types de base (issus du lexique), types présuppositionnels (transmettant les contraintes issues des présuppositions), types disjoints et fonctionnels. Les types peuvent faire l'objets d'opérations (tels le calculs, dans la hiérarchie de type, du plus petit majorant ou du plus grand minorant) et de quantifications. Les notions de sous-typage et d'application sont définies axiomatiquement.

Le lexique apporte des contraintes sur le type présuppositionnel des variables employées ; les constructions syntaxiques et le discours peuvent apporter des contraintes sur le contexte à la manière de la sémantique dynamique, en affirmant des relations sur les types des variables. Les définitions de ce système pourraient gagner en clarté et lisibilité, mais le résultat est un ensemble de contraintes de types détaillées argument par argument.

L'auteur souhaite utiliser des types de nature hyperintesionnelle, qui ne disposent pas d'interprétation ensemblistes. Selon lui, la logique des types de Martin-Löf pourrait donner un modèle satisfaisant, mais il préfère, pour donner une interprétation personnelle notamment des types complexes, se baser sur un modèle catégorique. Il commence par associer à chaque type une catégorie cartésienne fermée, dont les opérations intrinsèques sont un modèle cohérent pour le produit, la disjonction ou les types fonctionnels, puis lui ajoute une construction spécifique (les *objets associés*) correspondant à la notion de sous-typage. Enfin, après une longue discussion sur la nature intrinsèque des types pointés, il fait correspondre à cette construction un produit fibré asymétrique entre les catégories correspondant aux types le composant. Cette dernière construction est très délicate, et nous semble fort discutable ; une certaine confusion entre objets de la logique et modèles pour l'interprétation de ces derniers parasite les propositions.

Vision du monde
Au fil des constructions introduites et des discussions philosphiques, [Asher, 2011] établit une vision personnelle des modèles de la langue et du monde.

Notion de concepts
Les concepts de base renvoient aux lexèmes de la langue et aux types de la logique.

Ainsi, les concepts sont organisés en une hiérarchie ontologique représentant les connaissances du monde aussi bien que l'ensemble des mots de la langue. Les concepts les plus génériques sont au sommet de la hiérarchie, et leur raffinement correspond à des concepts de plus en plus élémentaires, jusqu'à décrire un élément unique.

Le système de types reflète cette hiérarchie de concepts. Même si le typage d'un élément peut sembler complexe, il ne fait que représenter les caractéristiques qui sont établies sur les concepts correspondant au moment de l'instanciation. Le fait qu'un symbole soit l'instance d'un type signifie simplement que le mot qu'il symbolise est porteur de l'information codée par le type.

Les contraintes sur les types, alors, représentent des relations entre concepts. Le fait qu'il existe une liaison analogique entre deux concepts sera représentée par une influence sur un type, qui devra être justifiée. Il y a également une notion d'ordre importante : l'information est d'abord apportée, puis traitée.

Il n'y a non pas une, mais deux logiques qui se superposent : la logique des concepts, qui permet de spécifier les types, leurs contraintes et relations, et la logique des éléments, qui constitue le calcul Montagovien de la forme logique de la phrase. Ces deux logiques sont jointes par la justification, opération qui consiste à opérer dans la forme de la phrase les modifications structurelles induites par celle du type qu'a reçu le terme. Symboliquement, le sens de la phrase est chargé de « justifier » l'information apportée au fil des constructions.

Applications
De très nombreuses constructions sont ainsi intégrées dans le canevas proposé par l'auteur.

Les aspects multiples des noms sont des conséquences directes de leur typage : un type pointé, composé de ces aspects duaux. Les noms peuvent recevoir des prédications simples, multiples, ou quantifiées suivant l'un ou l'autre des aspects.

La coercition est expliquée par des types dépendants. Cette relation de dépendance de types donne lieu, lors de sa « justification », à une modification de la prédication qui effectue la coercition.

La prédication restrictive (constructions avec *A vu comme B* ou *en considérant A en tant que B*) est perçue comme permettant l'introduction de types pointés, qui pourront permettre de considérer l'argument sous une facette particulière, puis une autre.

Les informations données dans le discours (par exemple, par référence anaphorique) ou les pluriels vagues sont traités au moyen de la pluralité : les types dépendants suffisent à justifier les phénomènes observés.

Certaines constructions syntaxiques, comme le génitif, apportent d'autres informations sur les types. Il s'agit de relations diverses, données par le lexique.

La nominalisation est vue comme une opération dans la logique de spécification des concepts, pouvant apporter une abstraction supplémentaire et donc changer l'ordre du type concerné.

Les mécanismes ordinaires de la logique et leur utilisation compositionnelle suffisent pour représenter les métonymies, sans qu'il y ait coercition.

De la même façon, plusieurs autres constructions (à commencer par les prédications simples) reçoivent leur représentation logique dans le système proposé.

Sémantique et réalité

La forme logique ainsi construite reflète la réalité, mais qu'en est-il des constructions non littérales ?

Dans l'optique de [Asher, 2011], le discours représente un état de fait sur les éléments du monde. Si l'auteur insiste pour employer des types fortement intensionnels (ainsi que des mécanismes hyperintensionnels), il fait constamment référence à une situation existante vérifiable, quasi-expérimentale. Les intentions servent surtout à représenter des situations hypothétiques, correspondant au conditionnel ou à des états de croyance.

Ainsi, la métaphore est représentée comme un discours à la signification littérale, mais qui introduit implicitement des types pointés ; ces types pointés apportent artificiellement une facette supplémentaire, sur laquelle porte la métaphore. Cette construction est analysée au niveau de la forme logique, et non de son interprétation.

Pour ce qui est des écrits de fiction, l'auteur emploie une approche radicale : les éléments fictifs son considérés comme une hiérarchie de sous-types des entités générales, dont aucun type n'admet d'instance. Les prédicats portent donc sur des interprétations sans aucun représentant extensionnel.

Enfin, au cœur du système proposé sont les mécanismes de transposition : qu'il s'agisse des types pointés, dépendants, disjoints, du sous-typage ou des relations lexicales, il s'agit toujours de transpositions (*maps*) d'un concept à un autre. L'auteur conçoit ces transpositions comme des objets de la métaphysique humaine, permettant de passer d'un concept à un autre ; s'il n'y a, dans aucune langue, une méthode pour opérer une coercition d'un type en un autre, c'est alors que la métaphysique humaine conçoit cette transposition comme impossible.

[Asher, 2011] considère donc que le langage est une opération de communication visant l'établissement de faits, en rapport direct avec la réalité.

Comparaison avec notre approche

Points forts par rapport à notre approche

[Asher, 2011] synthétise près de dix ans de travaux d'experts du domaine concerné ; par rapport au présent manuscrit, il a de nombreux points forts, notamment dans la conception et l'ensemble des phénomènes couverts.

Les présuppositions de types

Les présuppositions de types introduites par l'ouvrage sont novatrices, et n'ont pas d'équivalent dans notre approche.

En effet, bien que n'ayant que peu à voir avec les présuppositions traditionnellement étudiées en linguistique, elles permettent de contraindre l'information au travers d'un mécanisme portant exclusivement sur les types, là où nous opérons sur chaque variable. Les relations affirmées sur les types et leur justification forment un système puissant, générique, qui permet de ne donner des spécifications qu'au niveau de la hiérarchie des types. Notre approche opère sur les entrées lexicales individuelles, et est donc très loin de cette puissance.

La couverture des phénomènes

Plus importante encore, la couverture des phénomènes par l'ouvrage est sans commune mesure avec celle à laquelle nous pouvons prétendre.

L'accumulation d'exemples, et les années d'études que chacun a occasionné ont permis de rassembler une bibliothèque de phénomènes impressionnante, comportant des alternations syntaxiques, des constructions diverses, des exemples dans de multiples langues et des cas d'études ayant fait l'objet de nombreuses publications préalables. Et, par conséquent, le nombre de phénomènes inscrits dans le système logique présenté, et développés avec une solution détaillée, est bien plus important que le nôtre, ou que celui de toute autre étude depuis [Pustejovsky, 1995].

Points de convergence entre les deux approches
Par de nombreux aspects, le présent manuscrit et [Asher, 2011] se rejoignent.

Prédominence de la sémantique lexicale
Les deux approches sont basées sur d'une sémantique lexicale riche, comportant l'ensemble des informations nécessaires aux prédications de la langue ordinaire.

La tentation est grande de remiser les données que nous avons choisi de faire figurer dans la sémantique lexicale à la pragmatique, à l'interprétation ou à une obscure phase de réflexion supplémentaire, le locuteur se contentant, pour la sémantique de la phrase, d'opérer une analyse compositionnelle en liant suivant la syntaxe les seuls mots qu'il y trouve. Nous avons choisi de faire figurer ces informations dans le sens des mots car elles sont, pour nous, intrinsèques à chaque lexème ; de nombreux exemples montrent que le sens des phrases diffère lors de remplacement de mots qui auraient une sémantique « simple » identique. Nous considérons avoir justifié cette approche, et le fait que nous parvenions à élaborer des systèmes pour l'ensemble des phénomènes rencontrés tend à nous y conforter.

Logique, types, ontologie
Les deux approchent partagent également un même postulat comme fondement de leur système logique.

C'est le résultat d'un point de vue convergent sur les théories précédentes, et notamment le fait que chaque approche cherche à proposer un système logique complet pour les phénomènes traités par [Pustejovsky, 1995], tout en adoptant de nombreuses thèses philosophiques de celui-ci.

Ainsi, toutes deux sont fondées sur une ontologie de concepts raffinée bien au-delà des types Montagoviens habituels. Toutes deux se basent sur une forme logique contrainte par des types issus de cette hiérarchie, et cherchent à exposer des opérations et modes de calcul qui permettent d'adapter, de modifier la sémantique d'un mot en cas de conflit de types.

Points faibles par rapport à notre approche

Malgré le soin apporté lors de la rédaction de [Asher, 2011], plusieurs aspects, tant techniques que philosophiques, y restent criticables.

Points techniques

C'est d'abord le système logique qui pose problème.

Les types composites pour un même terme, les opérations multiples parfois très complexes, et les deux niveaux de logique (logique de spécification de type et forme logique de la phrase) sont à clarifier, ce qui peut conduire à des corrections futures.

De même, la présence de nombreux produits pour les types entraîne une certaine confusion, même si toutes les constructions sont abondamment justifiées.

Mais c'est surtout l'utilisation des catégories qui est problématique. Il ne nous apparaît pas clairement s'il s'agit de présenter un système logique dont les objets sont des catégories, ou une interprétation d'un système logique dans les catégories. De plus, l'argumentation développée pour justifier ces constructions nous paraît discutable. Enfin, la présentation des types pointés en tant que produit fibré transfini de catégories ne nous paraît pas utilisable en l'état.

Points philosophiques

Nous avons également des divergences de fond.

La principale porte sur le rapport direct que fait l'auteur entre la réalité et le langage. Pour nous, une ontologie de la langue ne correspond pas forcément à une ontologie de la réalité, de même qu'un acte de langage n'exprime pas forcément un point de vue sur la réalité, mais plutôt sur un monde construit, dépendant de l'énonciation, et qui peut être conçu selon une représentation minimale. Ainsi, nous pensons que les mécanismes d'inférences basés sur les types et un ensemble de transpositions donné par la métaphysique sont probablement trop généraux pour prévoir les expressions du langage, ce dernier comportant de nombreux idiomes.

En conclusion

L'approche de Asher éclipse la nôtre sur de nombreux points, notamment la généricité de son système de présuppositions de types et la couverture des phénomènes. Cependant, nous pensons que notre approche est plus simple dans sa réalisation et échappe à certains écueils. Les deux approches partagent les mêmes fondations.

Nous considérons que les points développés par les deux approches sont différents, mais qu'elles sont complémentaires dans une théorie générale de la modélisation de la sémantique lexicale et discursive.

En effet, la différence principale entre les deux théories développées est la suivante : celle de Asher permet de faire des inférences génériques à partir des types, tandis que la nôtre se concentre sur les apports idiosyncratiques de chaque entrée lexicale. On peut raisonnablement imaginer que la réalité de la composition sémantique de la phrase fait appel à ces deux types de mécanismes.

7.2.4 Cooper

[Cooper, 2007] propose une formalisation des types complexes par le biais de la logique des registres (*record types*). Cette logique, basée sur la théorie des types de Martin-Löf (voir [Nordström et al., 1990]), comporte à la fois une représentation de types arbitrairement et récursivement complexes et une notion de sous-typage forte. Dans cet article, Cooper introduit également la notion de quantificateurs généralisés dynamiques, et donne des propositions directes pour la conversion des types complexes du lexique génératif ; il démontre qu'un tel paradigme est capable de gérer la co-prédication et l'usage novateur des mots, pour le moins.

Cette publication n'a néanmoins pas été suivie de discussions ; elle n'examine pas, en particulier, le cas des co-prédication non-élégantes.

Nous considérons que notre approche est alternative à celle de Cooper, et que les deux peuvent présenter un intérêt dans le traitement des types complexes.

Discussion

Contrairement aux approches citées précédemment, nous disposons d'opérations fortement dépendantes du lexique, et non pas dérivées des types employés. Il s'agit d'un choix délibéré qui peut faire perdre en rigueur et en généralisation notre formalisme, mais qui permet de lui donner de bons espoirs dans l'adéquation plus stricte aux phénomènes observés, et une certaine forme de flexibilité.

Il reste cependant critiquable en ce que le mécanisme précis de sélection des transformations n'est pas explicite ; le libre choix est assumé, ce qui peut donner lieu à des implémentations d'une forte complexité. Nous verrons que, pragmatiquement, ce n'est pas le choix retenu.

Chapitre 8

Implémentation

Une implémentation complète de la présente proposition reste à réaliser ; elle suppose de nombreux efforts en termes de modélisation des opérations et de cadrage de leur portée, notamment au second ordre. Cependant, une implémentation partielle des principes développés ici a été effectuée, intégrant sous la forme d'un module de l'analyseur syntaxique et sémantique *Grail* les mécanismes d'adaptation sémantique des lexèmes basée sur des types raffinés, avec un certain succès.

8.1 Cadre de l'implémentation réalisée

Cette première implémentation est la réalisation associée à un stage de l'École Normale Supérieure – Cachan (antenne de Bretagne) en 2010, effectué sous la direction de M. Christian Retoré par M. Emeric Kien.

Au cours de ce stage effectué en étroite collaboration avec nos travaux de recherche, M. Kien a été amené à effectuer des études approfondies sur les bases du traitement automatique des langues, les formalismes d'analyse syntaxique et sémantique utilisées dans *Grail* (les grammaires catégorielles et la sémantique de Montague, entre autres), ainsi que sur la sémantique lexicale et nos propositions de système formel. Se fondant sur les implémentations modulaires de *Grail* déjà réalisées, il a ensuite développé un module destiné à l'implémentation de la plus grande partie de nos propositions, et a procédé à des essais sur un jeu de tests.

Ce travail a duré deux mois (un consacré aux études préliminaires et un à l'implémentation proprement dite), et il s'agit d'un stage de recherche ; il n'était donc pas question de donner une implémentation complète et achevée de l'ensemble des théories développées ici, mais plus de démontrer, au travers de cas concrets, la faisabilité d'une analyse sémantique basée sur les principes généraux de celles-ci.

8.2 L'analyseur syntaxique et sémantique *Grail*

Cette implémentation s'appuie sur un analyseur syntaxique et sémantique existant, *Grail*.

Développé par M. Richard Moot au sein de l'équipe SIGNES, *Grail* est un analyseur conçu pour les grammaires catégorielles multimodales. Au cours de son développement, il a été rendu compatible avec tous types de grammaires : grammaires de Lambek, de discontinuité, de Carpenter… mais également des grammaires plus linguistiques, comme HPSG.

Les grammaires sont données sous la forme d'un lexique et d'un ensemble de règles structurelles. Le lexique est considéré ici comme la donnée d'un mot (précisant l'entrée lexicale), d'une formule (précisant la syntaxe) et d'un λ-terme (précisant la sémantique) ; à partir de ces éléments, et étant donnée une phrase, l'analyseur calcule simultanément la syntaxe et la sémantique de cette dernière. Suivant la correspondance de Curry-Howard, *Grail* donne en sortie un λ-terme correspondant à la sémantique de l'analyse effectuée.

L'analyseur est implémenté en SWI-Prolog, et est constitué d'un ensemble de bibliothèques. Au cœur de celles-ci, un mécanisme de calculs fondés sur les réseaux de démonstration (voir [Moot, 2007]) : des fonctions de recherche de preuves et de vérifications, notamment. Les réseaux de démonstration constituent un modèle opérationnel complet aussi bien pour la syntaxe des grammaires catégorielle que pour la sémantique du λ-calcul typé.

D'un intérêt particulier pour ce stage, le système d'inférence de types appartenant à l'analyseur sémantique effectue des vérifications en prenant appui sur le terme spécifié dans le lexique, examinant la cohérence des types de chacun des sous-termes lors de l'analyse, et en s'assurant de leur utilisation.

Au fil du temps, *Grail* a été plusieurs fois ré-écrit. Il a bénéficié d'améliorations de performances et d'efficacité par l'ajout de stratégies de filtrage, de la construction de grammaires pour le français et le néerlandais ; un lexique sémantique, basé sur la *Discourse Representation Theory*, et un autre, basé sur la sémantique dynamique de Philippe de Groote, sont en cours de développement… L'analyseur est, de manière permanente, en cours d'amélioration. Dans ce cadre, une première implémentation du lexique génératif avait été effectuée, mais les résultats n'étant pas probants, ce module avait été abandonné.

8.3 Choix effectués

Au cours du stage, de nombreux choix ont été effectués afin d'obtenir une théorie simple, compacte, qui puisse faire l'objet d'une implémentation rapide, efficace et vérifiable.

Ainsi, l'ontologie de types a été conçue comme une arborescence qui peut être représentée par une simple relation de sous-typage, plutôt qu'un treillis doté de propriétés spécifiques. Il suffit donc de spécifier la relation de sous-typage pour décrire l'ontologie ; c'est chose faite dans le lexique associé à la grammaire.

Pour représenter les manipulations correspondant aux termes d'ordre supérieur, des variables de types ont été utilisées, et on dispose d'opérations, de contraintes et de vérifications sur ces variables particulières. De cette manière, la résolution par unification de variables de Prolog suffit à rendre compte de nos opérations les plus complexes ; la quantification n'est cependant pas traitée.

Dans notre modèle, de multiples possibilités d'interprétation peuvent s'offrir lors de l'analyse ; dans cette implémentation, il a été jugé qu'une solution unique était à retenir pour une même analyse. Ont donc été définies des priorités sur les opérations. En particulier, l'attribution de types utilise en priorité les types les plus spécifiques possibles.

Le cœur de notre proposition, les modificateurs sur les termes induits par le lexique, ont été modélisés par des relations portant sur les types (également définis dans le lexique de la grammaire). Ces modifications sont inférées directement en cas de relation d'hyponymie ou d'hyperonymie, et spécifiées explicitement dans les autres cas ; elles apparaissent comme des fonctions introduites par le mécanisme d'inférence de types lors de l'analyse.

La plus grande difficulté pour l'implémentation a été la transposition des « degrés de rigidité », indispensables pour éviter les co-prédications inélégantes, entre autres. Le choix effectué est celui d'une spécification de portée, indexée par un entier, qui permet de traiter la plus grande partie des cas en signalant une erreur pour les modifications non valides.

8.4 Implémentation effectuée

Le système d'inférence de types de Grail a été augmenté pour tenir compte des relations suivantes :

```
upper_type(X,Y) :- sub_type(X,Y).
upper_type(X,Y) :- sub_type(Z,Y), upper_type(X,Z).
```

Il est ainsi possible de gérer, via ces relations de sous-typage généralisées, la hiérarchie des types. On définit ainsi un moteur d'inférence de types, déterminant automatiquement les transformations à appliquer :

```
% infer (+Type to infer,+Type inferred,-Transformation,-Priority)
infer(X,X,[],0).        infer(X,Y,[],0) :- upper_type(Y,X).
infer(X,Y,[X,Y],P) :- may_infer(X,L),Y2=[Y,P],member(Y2,L).
infer(X,Y,A,P) :- upper_type(Z,X),infer(Z,Y,A,P).
infer(X,Y,A,P) :- upper_type(Y,Z),infer(X,Z,A,P).
infer(T1->T2,T3->T2,[],P) :- infer(T1,T3,[],P).
```

On ajoute également une structure permettant de spécifier les transformations et informations de types associées à chaque terme complexe :

```
% typed_sem(+LambdaTerm, -TypedLambdaTerm, -Type, +[],
      -List of transformations, +[], -Types of variables)
```

À partir de ces primitives, on peut ensuite définir un lexique-test, comprenant une hiérarchie de types définie très simplement :

```
sub_type(e, physical).
sub_type(physical, artifact).
sub_type(artifact, commodity).
sub_type(commodity,vehicle).
sub_type(vehicle, car).
sub_type(car, honda).
sub_type(commodity,dress).
sub_type(artifact, building).
sub_type(physical, natural).
sub_type(natural, person).
sub_type(natural, rock).
sub_type(e, abstract).
sub_type(abstract, thought).
sub_type(abstract, information).
sub_type(information, book).
sub_type(book,novel).
sub_type(abstract, institution).
sub_type(institution, bank).
sub_type(institution, city).
sub_type(city, capital).
sub_type(institution, government).
sub_type(institution, population).
sub_type(abstract,locus).
```

Nous pouvons également ajouter des inférences de type simulant le comportement souhaité de la polysémie pour des paires termes-types particulières :

```
may_infer(book,[[artifact,0]]).
may_infer(bank,[[building,1]]).
```

Les entrées lexicales seront alors relativement « classiques » :

```
entry(rock,n,rock,rock->t).
entry(car,n,car,car->t).
entry(honda,n,honda,honda->t).
```

Les définitions adaptées, et de multiples aménagements aux systèmes de résolutions internes de Grail, complètent cette implémentation.

8.5 Une preuve de fonctionnement

L'implémentation réalisée permet de vérifier que le modèle proposé peut, au moins partiellement, être validé.

Les ajouts les plus sensibles effectués sur *Grail* sont l'extension drastique du système de types (les types considérés habituellement par l'analyseur étant les types Montagoviens des entités et des valeurs de vérité), la prise en compte d'une notion de sous-typage, qui permet de spécifier une hiérarchie arbitraire de types sans modification du système d'inférence, et l'utilisation de variables de types. Ces opérations relativement simples permettent de spécifier une très grande partie des phénomènes à traiter.

Pour chaque étape, des vérifications de validité ont été effectuées sur des exemples concrets ; tous se sont révélés fonctionnels, avec une inférence instantanée (moins d'une dizaine d'étapes de calcul pour une phrase donnée).

Exemples de cas traités correctement :

- Le livre est posé sur la table.
- Je suis garé devant la maison.
- Washington est une ville cosmopolite à l'est des États-Unis.

Les exploitations de *qualia*, transferts de sens et co-prédications sous de multiples aspects sont ainsi traités, mais pas les quantifications.

En détail, voyons l'analyse effectuée par le programme de quelques cas…

8.5.1 Inférences simples

Étant donné l'exemple suivant :

```
example(" My honda is blue", s).
```

Le programme analyse la sémantique suivante :

```
appl(appl(lambda(A, lambda(B, appl(appl(is,
    appl(A, lambda(D, true))), B))), lambda(E,
        appl(lambda(F, lambda(G, bool(appl(F,
            G), &, appl(blue, G)))), E))), appl(my, honda)).
```

Après réduction, nous obtenons la forme logique suivante :

```
appl(appl(is, blue), appl(my, honda)).
```

L'analyseur traite de nombreux cas d'inférences, comme l'hyponymie (exemple du livre *Fight Club*, considéré en tant qu'artefact) :

```
example(" My fight_club is blue", s).
%
%% = Solution 1 =
%
%% = Semantics
%
appl(appl(lambda(A, lambda(B, appl(appl(is,
    appl(A, lambda(D, true))), B))), lambda(E,
        appl(lambda(F, lambda(G, bool(appl(F,
            G), &, appl(blue, G)))), E))), appl(my,
                fight_club)).%
%% = Reduced Semantics
%
appl(appl(is, blue), appl(my, inference(fight_club,
    ((book->t)->artifact->t)))).
```

Par défaut, l'analyseur considère que les phrases qui lui sont données sont sémantiquement justes. S'il ne connait pas d'inférence entre deux termes ou types, il supposera qu'une inférence inconnue appropriée existe ; à l'utilisateur de tirer les conclusions qui s'imposent.

```
example(" John wears a honda", s).
%
%% = Solution 1 =
%
%% = Semantics
%
appl(appl(wear, appl(a, honda)), john).
%
%% = Reduced Semantics
%
appl(appl(wear, appl(a, unknown_inference(
    honda, (dress->t)))), john).
```

8.5.2 Influence sur les prédicats

L'apport d'information par les prédicats est également gérée, permettant de traiter les exemples de Nunberg :

```
example(" John is parked", s).
%
%% = Solution 1 =
%
%% = Semantics
%
appl(appl(lambda(A, lambda(B, appl(appl(is,
    appl(A, lambda(D, true))), B))), lambda(E,
        appl(lambda(F, lambda(G, bool(appl(F, G),
            &, appl(parked, G)))), E))), john).
%
%% = Reduced Semantics
%
appl(inference(appl(is, parked),
    ((car->t)->person->t)), john).
```

Cette inférence présente dans le prédicat reste optionnelle, et ne gène pas l'analyse des prédications simples :

```
example(" A honda is parked", s).
%
%% = Solution 1 =
%
%% = Semantics
%
appl(appl(lambda(A, lambda(B, appl(appl(is, appl(A, lambda(D,
    true))), B))), lambda(E, appl(lambda(F, lambda(G,
        bool(appl(F, G), &, appl(parked, G)))), E))),
            appl(a, honda)).
%
%% = Reduced Semantics
%
appl(appl(is, parked), appl(a, honda)).
```

De plus, il reste possible de combiner d'autres prédications une fois cette inférence prédicative effectuée :

```
example(" A parked person is happy", s).
%
%% = Solution 1 =
%
%% = Semantics
%
appl(appl(lambda(A, lambda(B, appl(appl(is, appl(A,
    lambda(D, true))), B))), lambda(E, appl(lambda(F,
        lambda(G, bool(appl(F, G), &, appl(happy, G)))),
            E))), appl(a, appl(lambda(I, lambda(J, bool(appl(I,
                J), &, appl(parked, J)))), person))).
%
%% = Reduced Semantics
%
appl(appl(is, happy), appl(a, lambda(type(J, person),
    bool(appl(person, J), &, appl(inference(parked,
        ((car->t)->person->t)), J))))).
```

8.5.3 Inélégance des co-prédications

Comme demandé, l'utilisation de co-prédication inélégantes (pour une ville, portant à la fois sur la politique du gouvernement et sa population, par exemple) est détectée : l'analyse est effectuée normalement, mais le conflit entre aspects est mis en évidence et le programme indique une erreur.

```
example(" Paris is cosmopolitan and peaceful", s).
%
%% = Solution 1 =
%
%% = Semantics
%
appl(appl(lambda(A, lambda(B, appl(appl(is, appl(A, lambda(D,
    true))), B))), lambda(E, appl(appl(appl(lambda(F,
        lambda(G, lambda(H, appl(G, appl(F, H))))),
            lambda(I, appl(lambda(J, lambda(K,
                bool(appl(J, K), &, appl(peaceful,
                    K)))), I))), lambda(L, appl(lambda(M,
                        lambda(N, bool(appl(M, N), &,
                            appl(cosmopolitan, N)))), L))),
                                E))), paris).
%\\ %% = Reduced Semantics
%
error(appl(appl(is, lambda(type(N, capital), bool(appl(peaceful,
    inference(N, (capital->government))), &, appl(cosmopolitan,
        inference(N, (city->population)))))), paris)).
```

8.5.4 Groupes nominaux

L'analyseur gère aussi les prédications dans les groupes nominaux, qu'elles soient simples :

```
example(" My blue honda", np).
%
%% = Solution 1 =
%
%% = Semantics
%
appl(appl(lambda(A, lambda(B, appl(appl(is, appl(A, lambda(D,
    true))), B))), lambda(E, appl(lambda(F, lambda(G,
        bool(appl(F, G), &, appl(blue, G)))), E))),
            appl(my, honda)). %
%% = Reduced Semantics
%
appl(appl(is, blue), appl(my, honda)).
```

Ou complexes (avec une inférence inconnue, ici) :

```
example(" My blue idea", np).
%
%% = Solution 1 =
%
%% = Semantics
%
appl(appl(lambda(A, lambda(B, appl(appl(is, appl(A, lambda(D,
    true))), B))), lambda(E, appl(lambda(F, lambda(G, bool(appl(F,
        G), &, appl(blue, G)))), E))), appl(my, idea)).
%
%% = Reduced Semantics
%
appl(appl(is, unknown_inference(blue, (thought->t))),
    appl(my, idea)).
```

8.5.5 Article défini

Enfin, des débuts de quantifications ont été introduites pour la gestion de l'article défini ; cette quantification-ι ne pose pas les mêmes problèmes que les quantificateurs existenciels, universels, ou généralisés qui manquent toujours.

```
example(" The blue honda", np). %% = Solution 1 = %
%% = Semantics
%
appl(lambda(A, quant(iota, B, appl(A, B))), appl(lambda(C, lambda(D,
    bool(appl(C, D), &, appl(blue, D)))), honda)).
%
%% = Reduced Semantics
%
quant(iota, type(B, honda), bool(appl(honda, B), &, appl(blue, B))).
```

8.6 Le futur de l'implémentation

Si ce stage a permis de mettre en évidence la faisabilité d'une implémentation complète de nos propositions, cette dernière reste encore à réaliser.

Une première étape consisterait en l'amélioration du module présenté ici. Sans limite de temps et de moyens, il serait relativement aisé de prendre en compte des phénomènes additionels (comme les quantifications) et de raffiner les mécanismes existants afin de proposer une gestion de la rigidité plus réaliste ou une multiplicité des interprétations données. Il pourrait alors être nécessaire d'apporter des modifications ponctuelles au système d'inférence de types de *Grail*, en veillant à garder une complexité raisonnable pour les exécutions.

Mais une implémentation véritablement achevée de nos propositions devrait probablement être découpée en programmes indépendants (s'appuyant ou non sur un analyseur existant) ; l'utilisation cohérente de la logique du second ordre est difficile à envisager dans un autre cadre. En plus d'un analyseur donnant la forme logique d'une phrase, il serait souhaitable de proposer une interface permettant à l'utilisateur de définir les multiples lexèmes et transformations associées ; un module d'acquisition semi-automatique de données sera sans doute nécessaire à l'établissement d'un lexique sémantiquement riche pour une langue donnée.

Troisième partie

Prospective

Chapitre 9

Discussions et propositions

Les travaux présentés ici ne peuvent qu'avoir une couverture restreinte. Dans ce chapitre, nous proposerons de multiples directions, sans être exhaustifs, vers lesquels ils pourraient s'orienter dans l'avenir ; quelques systèmes à l'ambition très forte seront esquissés.

9.1 Couverture additionelle

Sans augmenter de quelque façon que ce soit les mécanismes proposés ici, nous pourrions examiner plusieurs extensions à la couverture phénoménologique de l'ensemble.

9.1.1 Agents implicites et autres phénomènes

Des recherches en cours (de Christian Retoré et Richard Moot, comportant un article soumis avec Laurent Prévot) portent sur les cas suivants :

Cinq heures de marches nous conduisent à Bordeaux. Cette longue route conduit à Bordeaux.

Ces exemples, et de nombreuses phrases similaires, portent sur le déplacement et la description d'itinéraires. Il apparaît que les verbes indiquant une destination, entre autres (*mener*, *conduire*, *aboutir*…) utilisent indifféremment des arguments divers : durée (du trajet), cheminement, moyen de transport. L'hypothèse retenue est que tous ces arguments font appel à un même mécanisme : la description partielle des actions d'un *agent implicite*, ici un humain, un voyageur qui effectuerai le parcours (pendant une certaine durée, suivant un chemin donné, ou par un certain moyen) ; dans tous les cas, le prédicat précise alors la destination de cet agent implicite.

Notre approche permet facilement de traiter ces cas, en supposant que les prédicats concernés portent tous sur un agent. Le lexique doit alors préciser les transformations des divers arguments permettant d'y associer ce voyageur virtuel.

Notons que cet agent implicite peut être utilisé dans d'autres situations, et peut se rapprocher de la notion d'*agentif* de Pustejosky, sans y être confondue (ce rôle étant restreint aux personnes étant à l'origine de l'action).

9.1.2 Proximité et associations des concepts

Une problématique qui peut être examinée à la lumière de notre proposition est celle de la validité sémantique de certaines alternations, en particulier l'utilisation du pronom personnel en opposition avec celle de l'article défini. Voyons quelques exemples :

Ce film était long, mais la mise en scène était intéressante.
Ce film était long, mais sa mise en scène était intéressante.

Ce concert était exceptionnel. Le soliste était divin.
? Ce concert était exceptionnel. Son soliste était divin.

Ce match était plein de surprises, l'arbitre était incompétent.
? ? Ce match était plein de surprises, son arbitre était incompétent.

Le repas était bon. Le dessert était particulièrement réussi.
? ? ? Le repas était bon. Son dessert était particulièrement réussi.

Dans tous les cas examinés ici, l'argument est un terme présentant plusieurs facettes, dont un événement. La gestion du lexique génératif est celle des objets complexes : les deux aspects sont mis sur le même plan, et il n'est pas possible de faire état de la relative inélégance de certaines phrases employant le pronom personnel.

Notre analyse est la suivante : les termes employés ici sont, avant tout, des événements. Ils disposent de multiples aspects très divers, mais tous ne sont pas associés aussi fortement à l'événement concerné. Pour chacun des événements, certains aspects sont directement accessibles, car ils y sont fortement liés : associés à cet événement en particulier. Pour ces aspects, il est possible de faire des références directes, à l'aide du pronom personnel.

D'autres aspects ne sont pas liés à l'événement de cette manière, car ils existent indépendamment de celui-ci : le soliste est toujours soliste après le concert, de même, l'arbitre a une vie en-dehors du match, et un plat, une entrée, un dessert peuvent être servis en-dehors d'un repas précis, ayant une date, un début, une fin. Pour ces aspects, il est plus difficile et moins élégant de faire une référence au moyen du pronom personnel.

Enfin, l'article défini permet de faire des références non pas à un terme précis, mais au contexte induit par l'emploi de ce terme. Il devient alors possible d'associer des concepts qui ne sont que très anecdotiques pour le terme employé, et qui ne peuvent raisonnablement faire partie de l'information lexicale. Par exemple : *Ce repas était long. L'entrée était à peine chaude. Le serveur était aimable, mais la ventilation était capricieuse.*

Cette réflexion peut nous permettre de différencier les prédications qui font appel à l'information lexicale de celles modifiant le contexte de l'énonciation.

9.2 Questions supplémentaires

9.2.1 Sens immédiat et double-sens

Le fait de pouvoir dégager de multiples interprétations possibles à partir d'une même phrase peut, dans une certaine mesure, permettre de mettre en évidence un effet de style permettant un double sens dans cette phrase. La comparaison entre les représentations données par ces interprétations peut être un indice important de ce cas.

9.2.2 Particularités d'une langue donnée

Pour faire apparaître les constructions idiomatiques propres à une langue ou un vocable donné, il suffit, dans notre construction, de faire apparaître un filtre fondamental différent pour chaque langue. Cela ne permettra pas de résoudre les problèmes de traduction inhérents à ces cas précis.

9.2.3 Acquisition, apprentissage

Sans ajouter de nouveaux mécanismes à ceux proposés ici, on peut proposer une méthodologie simple pour l'enrichissement semi-dynamique d'un lexique existant, et donc l'acquisition de nouveaux lexèmes et transformations : considérer toute phrase comme valide, et ajouter dynamiquement les termes manquants au filtre en usage. Ce dernier pourra ensuite être validé par un locuteur compétent.

9.3 Mécanismes de la représentation des connaissances pour la sémantique

9.3.1 Lexique et couches de filtres

Nous allons maintenant proposer une extension à notre canevas permettant le traitement de différents contextes.

Pour gagner en facilité d'utilisation du mécanisme, l'architecture lexicale sera décrite en *couches*, la couche la plus haute étant utilisée, si le mot y est présent, avant chacune des couches inférieures, ce qui permet de *surcharger* des opérations, par exemple.

Le lexique fondamental

La couche la plus basse est celui du lexique de la *langue*, c'est-à-dire d'un dictionnaire correspondant au langage couramment admis comme de référence. Il comporte de très nombreuses entrées, certaines polymorphes, et est le plus souvent utilisé.

Univers locaux

Dans le cadre d'un texte particulier, tel que livre ou série de livres, on pourra ajouter une couche correspondant au vocabulaire particulier à l'*univers* employé : s'il s'agit de fiction, notamment de conte, *fantasy* ou sience-fiction, en particulier, mais également dans le cadre d'un ouvrage technique d'un certain domaine. Ces couches d'univers, ou de domaine, ont un certain nombre de mots, peuvent s'enrichir par hypostase ou définition (ainsi, l'univers de *Star Wars* comporte des mots définis par morphologie du français tels que *transparacier, plastacier, sabre-laser…*), et on y accède fréquemment.

Filtres de petite taille

Enfin, le discours peut introduire des *filtres*, ou couches de petite taille, situées au-dessus des autres. Elles correspondent à des situations langagières particulières, comme *Parlons maintenant de Jean, en tant que personne/animal/banquier/...*, *Imaginons un instant que les vaches ailées existent*, ou autres actes de parole. Ce sont des lexiques comportant quelques mots et quelques opérations, utilisés sur le moment et dont on se débarasse ensuite.

9.3.2 Fonctionnement des filtres

On parlera désormais du *lexique* de la langue en tant que couche la plus profonde, toujours active, et de *filtres* plus superficiels d'univers ou de discours.

Activation d'un filtre

Au début de l'analyse d'un texte, on se place dans le lexique de référence, et on crée un filtre d'univers, *a priori* vide sauf si l'univers ou le domaine du texte est déjà connu. Par la suite, on active un filtre superficiel au repérage d'un mot présent dans ce filtre.

Modification d'un filtre

Le filtre universel en cours d'utilisation peut être modifié par définition ou hypostase, c'est-à-dire quand un nouveau lexème est rencontré, et qu'il ne peut être traité autrement. On peut également le modifier en cas de conflit de types, si aucune interprétation opérationelle ne peut être effectuée, en ajoutant à un lexème une opération *ad hoc*, vide de sens jusqu'à plus ample information.

9.3.3 Création à la volée d'un filtre

Les actes langagiers créatifs permettent de créer un filtre superficiel, contenant très peu d'informations. Les conditions d'entrée et de sortie de ce filtre sont comparables, et symétriques, à l'entrée / sortie d'une DRT.

9.3.4 Microcosmes : vers une représentation multi-agents

La notion que nous allons introduire ici ne cherche pas à se substituer à une éventuelle étape de « raisonnement » sur la sémantique, ou à permettre au système analystique de le « comprendre ». Il s'agit simplement de prendre en compte des éléments contextuels supplémentaires : quel agent a exprimé le texte ? Quelles sont ses connaissances ? Quelles sont les entités mises en jeu dans le discours, quels sont les faits établis sur ces dernières ? Y a-t-il des particularités lexicales associées à cet agent ?

Certaines de ces questions, comme beaucoup en sémantique lexicale, ont été évacuées comme relevant de la « pragmatique ». Cependant, il nous apparaît probable qu'elles participent de la dynamique des interactions sémantiques ou discursives – ou, du moins, qu'elles fassent partie des questions que nous pouvons traiter avant d'avoir recours à des modules de plus haut niveau.

Définitions

Entités
Un terme symbolique normal de ΛTY_n, non réductible à une valeur de vérité, est désignée comme *entité*.

Faits
Un prédicat associé à une entité, ayant un type vériconditionnel, est un *fait*. L'ensemble peut, naturellement, former une formule complexe (les prédicats étant également des entités).

Microcosme
Un ensemble d'entités et de faits portant sur ces entités. La définition est récursive : des microcosmes peuvent faire partie d'un microcosme, et des faits sur ces microcosmes également ; on appelle alors ces derniers *jugements*.

Hiérarchie

Le principe de fonctionnement de ces catégories est de créer, pour chaque phrase analysée, un microcosme élémentaire ne contenant que les formules logiques directement associées à cette phrase. Par défaut, la valeur associée est *vrai* : lors de l'analyse d'une phrase, le locuteur imagine d'abord un monde dans lequel la situation présentée est vraie (le microcosme), puis le confronte à son expérience afin de décider de sa pertinence (intègre au microcosme de son expérience le microcosme nouvellement produit avec un jugement tel que « vrai », « faux », ou « correspondant à une situation », laquelle peut être réelle, circonstanciée, ou fictive[1]).

On dispose donc d'une hiérarchie simple : un microcosme représentant le « savoir » d'un agent, composé de microcosmes étiquetés de jugements formant plusieurs catégories : les microcosmes associés au jugement *vrai* représentent les connaissances applicables à la « réalité » de l'agent, ceux associés à un symbole précis, aux connaissances de l'agent sur tel ou tel domaine fictif (roman, film, mythe…).

Introduisons, en plus du microcosme représentant le « savoir », les « connaissances », ou « l'expérience » (au choix) de l'agent, un second microcosme représentant le « lexique ». Ce dernier est composé de faits portant sur des entités de bases qui sont des mots, et contenant l'intégralité des informations nécessaires à l'analyse de la sémantique lexicale. Le lexique peut être hiérarchisé en plusieurs microcosmes, différenciant les langues (ou dialectes) employés, tout d'abord, ainsi que les autres « filtres »établis au-dessus du lexique (voir précédemment – un filtre peut s'appliquer dans le cas d'un univers fictif déterminé, par exemple). Ce sont ces informations qui servent de base à l'analyse sémantique des phrases perçues par l'agent, et donc à la génération des microcosmes de son « savoir ».

Principes et mise en œuvre des agents

Un agent est donc un microcosme, comprenant au moins un microcosme « savoir » et un microcosme « lexique ». Pourquoi hiérarchiser davantage ? Pour prendre en compte les possibilités du récit et du dialogue (à la manière des récents développements dans ce dernier domaine), il faut pouvoir différencier chacun des agents apparaissant, que ce soit directement (acteurs du dialogue) ou non (référencés dans le dialogue ou récit).

[1] Il est nécessaire de tenir compte de l'expérience de la fiction pour permettre de construire un agent se représentant simultanément plusieurs romans, ou acquérant simplement un corpus sans contrôle de ce dernier.

Il est intéressant de voir que récit et dialogue peuvent être infiniment hiérarchisés (la profondeur de cette hiérarchie ne dépend que de la performance du locuteur) : un récit peut contenir un dialogue, dans lequel un des acteurs fait référence à un autre agent, etc.

Dans ce cas, la hiérarchie prend tout son sens : lors de l'analyse d'une phrase prononcée par un agent A, analyse effectuée par un autre agent B, le microcosme associé à cette phrase aura pour jugement l'identifiant de A. D'autre part, un microcosme représentant l'état des connaissances de B sur A sera mis à jour ; il servira, pour B, à formuler les inférences qui seraient faites par A. De cette façon, une représentation minimale (et potentiellement faussée) d'un agent peut être contenue dans l'ensemble des connaissances d'un autre, afin de permettre à un agent d'adopter, du mieux qu'il le puisse, son point de vue.

Il ne s'agit pas ici de faire du raisonnement de haut niveau, simplement d'inférer la valeur de vérité des faits en utilisant d'autres prémisses.

Dans le cas du récit, le monde imaginé par le lecteur comprend l'ensemble des points de vue des différents agents auxquels il est fait référence : c'est pour correspondre à cette intuition que ce système a été conçu.

Chapitre 10

Conclusion

Au cours des années passées, nous avons été amenés à nous familiariser avec des domaines étrangers, mais pourtant au cœur de la théorie de l'information : les formalisations de l'analyse sémantique et du lexique. Des études approfondies nous ont permis d'assimiler l'histoire de ces concepts et les récents développements de ces derniers, et nous ont conduit à nous interroger sur plusieurs théories récentes. De nombreuses publications prouvent que la recherche en ce domaine est toujours active, et se poursuivait en parallèle de nos travaux. Forts de ces références, nous avons cherché à développer un cadre d'intégration pour un lexique à informations riches dans la sémantique de Montague, avec l'ambition de l'intégrer aux outils déjà développés dans ce domaine.

Au fil de ces études, un ressenti restait constant : chaque publication se montrait insatisfaite des résultats des travaux l'ayant précédée, dans ce domaine précis bien plus qu'ailleurs. C'est cette insatisfaction que nous retenons le plus facilement, par rapport à l'ensemble des travaux de si nombreux et éminents experts, mais également par rapport à nos propres apports, bien modestes en regard du chemin parcouru. C'est que la sémantique lexicale est un domaine difficile, qui réclame des efforts pouvant paraître insurmontables ; la formaliser entièrement, de façon satisfaisante, reviendrait à avoir donné une modélisation des phénomènes langagiers les plus divers et les moins formels.

La théorie que nous avons présenté au sein de ce manuscrit n'a pas pour prétention de répondre à l'ensemble des questions posées par la sémantique lexicale, mais de donner un cadre capable de relier les différents collaborateurs de cette couche de la langue, en particulier syntaxe et sémantique. En reprenant les types raffinés proposés par le lexique génératif et leur hiérarchie d'héritage ontologique, en donnant une analyse sémantique proche de Montague pour la phrase tout en permettant, par l'utilisation d'une logique du second ordre, l'introduction de transformations données par le lexique, nous pensons avoir proposé un cadre d'analyse et de calculs qui permet de traiter avec satisfaction la plupart des phénomènes rencontrés au niveau de la phrase ; nous avons également évoqué des raffinements à ce canevas qui pourraient permettre la prise en compte du contexte au-delà de la phrase, et même du texte, en intégrant les différences entre les lexiques des différents agents impliqués.

Pourtant, ce canevas manque de généricité et nécessite des mécanismes puissants, ainsi qu'un lexique très travaillé, afin d'être mis en place. De nombreuses objections restent en suspens, et la théorie présentée demeure imparfaite ; c'est inévitable dans le cadre des travaux réalisés.

Nous avons également évoqué l'implémentation réalisée d'une partie de notre proposition, effectuée en tant que module d'un analyseur syntaxique et sémantique déjà existant. Elle permet de démontrer la faisabilité d'un programme donnant des résultats concrets à partir d'une analyse basée sur la sémantique lexicale.

Cependant, cette implémentation reste largement inachevée.

Lorsque nous avons commencé à traiter le sujet, nombreux étaient les espoirs associés à sa résolution. Un ensemble de programmes traitant les informations sémantiques apportées par le lexique, l'intégration de la sous-spécification, l'analyse de phrases à sens multiples, la mise en évidence de certaines figures de style… Si nous sommes encore bien loin de ces multiples applications, nous considérons avoir progressé vers leur réalisation.

Il reste cependant de nombreuses études à effectuer. L'utilisation de la sémantique dynamique serait un premier pas, tout comme le serait une implémentation plus précise et complète. L'acquisition automatique de données pour le lexique paraît indispensable à moyen terme, et toute implémentation devra s'accompagner d'une recherche de précision et d'efficacité absente des présents travaux.

Il est également probable que le domaine particulier de la sémantique lexicale puisse bénéficier d'apports en provenance de la cognitique, car la compréhension du processus d'interprétation de la phrase serait primordiale pour donner des mécanismes d'analyse plus proches de la réalité.

Mais malgré tous les manquements, nous espérons que les présents travaux puissent servir en l'état pour une analyse précise de la sémantique des phrases. On pourra également les utiliser comme source d'exemples de phénomènes polysémiques, et comme référence sur les nombreux travaux qui nous ont précédé, et se poursuivent encore à l'heure actuelle.

Tout reste à faire ; nous avançons pas à pas vers la formalisation de la langue, mais l'objectif est encore lointain.

Mots-clés et références

ACG Pour *Abstract Categorial Grammars* : grammaires catégorielles abstraites. Il s'agit d'un formalisme modulaire, basé sur le parallèlisme entre les structures de la langue au niveau phénogrammatical, tectogrammatical et sémantique, d'une part, et les termes du λ-calcul, d'autre part. Voir [de Groote, 2001] et les développements plus récents, dont [De Groote et al., 2009].

λ-calcul Le λ-calcul est un système de représentation symbolique et de description formelle. Il est conçu et orienté pour la représentation de la notion de *fonction*, et permet de décrire de très nombreux phénomènes ; il peut être générique (non typé) ou restreint (typé). Dans ce manuscrit, nous employons le λ-calcul typé au second ordre (c'est-à-dire comprenant des variables de types), avec des notations classiques (application préfixée, abstraction, équivalence $\beta\eta$) ; voir par exemple [Girard et al., 1989].

λ-terme Un λ-terme est un terme du λ-calcul, c'est-à-dire un élément qui peut être utilisé pour construire un autre terme. Les termes sont définis par induction structurelles à partir des constantes du système et de mécanismes d'abstraction (λ), application, réduction (β), renommage (α), conversion (η), etc. Dans le λ-calcul typé, chaque terme dispose d'un type bien défini ; à l'ordre supérieur, ce type peut dépendre des opérations effectuées et de l'abstraction du second ordre (Λ).

Lexique Un lexique est un ensemble de définitions pour le vocabulaire d'un dialecte donné. Dans le cadre de l'analyse des langues, le lexique regroupe l'ensemble des informations associées par le locuteur à chaque *mot* (appelé *lexème*), indépendamment du contexte de son énonciation. Selon les approches, le lexique est plus ou moins étendu : dans de nombreux modèles Montagoviens, seules des informations de syntaxe (place attendue dans la phrase, accords…) et de sémantique élémentaire (type syntactique, arguments) sont données par le lexique. Dans le cadre de la sémantique lexicale, du Lexique Génératif et de ce manuscrit en particulier, le lexique est riche et comporte de très nombreuses informations : idéalement, tout ce dont le locuteur du dialecte a besoin pour l'analyse du sens précis de la phrase. En particulier, les mots ayant plusieurs sens, chaque lexème contient des informations sur son type sémantique, les relations et transformations associées, afin de pouvoir déterminer dans le contexte de toute énonciation le sens sous lequel il est employé.

Microcosme Au sens employé ici, un microcosme est la représentation minimale d'un monde partiellement défini. Il contient les représentations symboliques d'entités et de faits saillants portant sur ces entités. Dans la théorie esquissée dans les perspectives, nous proposons d'employer de telles représentations pour conserver un lexique local à un contexte donné par un couple agent / univers.

Modèle Un modèle est un paradigme d'interprétation. Dans l'analyse Montagovienne et ses successeurs, une phrase subit une analyse syntaxique puis sémantique, ce qui a pour résultat une formule logique symbolisant le sens de la phrase suivant ces analyses. L'interprétation de cette formule dans un modèle permet d'associer des objets mathématiques précis à la formule considérée, donc à la phrase. Un modèle possible est celui de la théorie des ensembles : les variables portent sur les individus, les prédicats définissent des ensembles (ceux des individus pour lesquels le prédicat est vrai), etc. Il est possible de raffiner ces modèles pour prendre de multiples phénomènes en compte. Notre réflexion ne porte *pas* sur les modèles de l'interprétation, mais plus sur les inférences possibles à partir des formules logiques.

Modélisation La modélisation d'un phénomène physique consiste en l'extraction de données pertinentes et leur synthèse sous une forme symbolique, plus facilement appréhendable que le phénomène original : on représente ainsi les forces par des vecteurs, les durées par des intervalles, etc. Dans le cas de la langue, le phénomène modélisé est constitué d'un corpus de phrases évoluant sans cesse. Les modélisations varient selon les philosophies du langage, certains étant d'avis que la langue est elle-même sa meilleure modélisation possible ; d'autres cherchent à modéliser certains de ses aspects : morphologie, syntaxe, sémantique, pragmatique, évolution…

Ontologie L'Ontologie, au sens strict, est un domaine d'études de la métaphysique et de la philosophie portant sur la description de l'existence : quelles entités existent, quelles sont leurs propriétés, comment les classer. Par extension, ce domaine a été transcrit dans l'informatique et les sciences de l'information, et on parle désormais d'ontologies. *Une* ontologie est un mode de description d'un ensemble, qui y associe généralement une hiérarchie et une taxonomie. Elle peut être générale (auquel cas elle devrait porter sur l'ensemble du réel) ou restreinte (ne concernant qu'un domaine précis), et, par extension, il peut être fait état d'une ontologie portant sur des concepts et non pas des entités réelles. C'est le cas ici, où l'ontologie considérée porte sur l'ensemble des mots de la langue, et y associe une hiérarchie hyponymique, classant du concept le plus général au terme le plus précis.

Réalité Dans un travail scientifique de modélisation, la réalité est prise comme point de référence. Cependant, ici, la modélisation porte sur des actes de langage ; il est nécessaire de nuancer l'importance de la réalité vis-à-vis des phénomènes ciblés. Les cas d'études sont des énonciations qui sont avérées (le simple fait de les transcire en donne un exemplaire), mais qui peuvent ne pas être fondées elles-mêmes sur des faits réels ; la plupart des exemples donnés ici sont des « phrases-jouet ». Il est donc difficile de donner des jeux de tests falsifiables, et nous sommes souvent amenés à utiliser une notion très subjective d'« élégance ».

Second Ordre Deux notions (pour le moins), très différentes mathématiquement, sont recouvertes par l'appellation *second ordre*. Dans le cas d'une *logique*, le second ordre signifie un ordre d'abstraction supérieur dans lequel on raisonne, non plus sur des individus, mais sur des ensembles arbitraires. Dans le cas qui nous occupe dans le présent manuscrit, nous parlerons de calculs et de termes *typés* au second ordre : il s'agit ici de permettre d'abstraire des *types* aussi bien que des termes. Un λ-calcul typé au second ordre présente donc un degré d'abstraction supérieur au λ-calcul typé au premier ordre, mais il reste plus restreint que le λ-calcul non typé.

Sémantique La sémantique recouvre un grand nombre de notions associées au *sens*. Ainsi, un lexique doit-il fournir la sémantique d'un mot, c'est-à-dire sa signification. Dans le cas de l'analyse sémantique formelle de la phrase, il s'agit d'un mode de calcul, supposant que chaque lexème a été préalablement associé à un terme contenant, entre autres, les directives à employer pour le combiner à d'autres termes ; l'analyse consiste alors en la réalisation de la combinaison de l'ensemble des sémantiques des mots composant une phrase pour aboutir à une représentation sémantique complète (en général, une forme logique) de cette dernière. La sémantique de chaque lexème comprend également un ou des symboles, permettant l'interprétation de cette sémantique suivant un modèle.

Sens Qu'est-ce que le sens d'un mot, d'une phrase ? Il ne nous appartient pas de donner une réponse exacte à cette question, si tant est qu'elle existe ; dans le cas de l'exploitation informatique de données langagières, nous nous contenterons d'associer les sens à une représentation symbolique, manipulable par un programme ou un utilisateur, afin de pouvoir opérer sur celle-ci diverses opérations.

TY_n Voir [Muskens, 1996]. Cette logique des types à n sortes est une extension du calcul Montagovien original, et de la logique intuitionniste. Sa seule originalité est que, en lieu et place des seules *entités*, nous pouvons disposer de n sortes distinctes ; ces sortes, avec le type t des valeurs de vérités, forment le lexique des types atomiques de la logique. TY_n est couramment utilisée pour des extensions aux principes Montagoviens, comme l'intensionnalité, l'hyperintensionnalité… Nous l'employons pour modéliser les types raffinés, ontologiques, introduits par le lexique génératif.

ΛTY_n Nous avons introduit ΛTY_n, un λ-calcul typé au second ordre basé sur TY_n. Mis à part le langage de types issu directement de la logique des types à n sortes, ce calcul est caclqué sur le Système F. Avec les modifications de l'application et les notions associées, ce calcul constitue une proposition complète de système formel pour le traitement de la sémantique lexical dans un cadre compositionnel, et constitue l'objet de l'essentiel de ce manuscrit.

Type Donner un *type* à un terme du λ-calcul sert de multiples buts. Un système de types élémentaire permet de donner le nombre d'arguments d'un prédicat, par exemple. Les types des langages de programmations fortement typés ont une autre raison d'être : ils précisent l'usage et la sémantique de chaque terme employé. Dans le lexique génératif et dans le présent manuscrit, nous employons les types pour imposer une classification sur les lexèmes, et donner des contraintes fortes à la prédication : ainsi, seuls les arguments de types *nourriture* pourront être utilisés pour le prédicat *manger*. Ces contraintes fortes, situées au sein même du système de calcul, permettent de définir de multiples opérations sémantiques dans ce même calcul : nous pouvons ainsi déduire de l'emploi du prédicat *manger* une tendance à modifier l'argument pour que son type devienne compatible avec celui de *nourriture*. La nature, l'emploi et les modalités d'applications de ces opérations, au sein de ce système de types raffinés par leur signification lexicale, constitue le cœur de ce manuscrit.

Références

[Anderson, 2006] Anderson, J. M. (2006). *Modern Grammars of Case*. Oxford University Press.

[Aristote, 350] Aristote (v. -350). *De la Physique, Livre II*.

[Asher, 2006] Asher, N. (2006). A Type Driven Theory of Predication with Complex Types. Unpublished.

[Asher, 2008] Asher, N. (2008). A Type Driven Theory of Predication with Complex Types. *Fundamenta Informaticæ*, 84(2) :151–183.

[Asher, 2011] Asher, N. (2011). *Lexical Meaning in Context : a Web of Words*. Cambridge University Press.

[Asher and Pustejovsky, 2005] Asher, N. and Pustejovsky, J. (2005). Word Meaning and Commonsense Metaphysics. Semantics Archive.

[Bally and Sechehaye, 1916] Bally, C. and Sechehaye, A. (1916). Cours de linguistique générale. D'après Ferdinand de Saussure.

[Bassac, 2006] Bassac, C. (2006). Morphologie et information lexicale. Habilitation à Diriger les Recherches.

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

[Blutner, 2002] Blutner, R. (2002). Lexical Semantics and Pragmatics. *Linguistische Berichte*.

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

[Carpenter, 1992] Carpenter, B. (1992). *The Logic of Typed Feature Structures*. Cambridge University Press, Cambridge, England.

[Chandrasekaran et al., 1999] Chandrasekaran, B., Josephson, J. R., and Benjamins, V. R. (1999). What are ontologies, and why do we need them ? *IEEE Intelligent Systems*, 14(1) :20–26.

[Chomsky, 1955] Chomsky, N. (1955). Logical Syntax and Semantics : Their Linguistic Relevance. *Language*, 31 :36–45.

[Chomsky, 1957] Chomsky, N. (1957). *Syntactic Structure*. Mouton.

[Chomsky, 1970] Chomsky, N. (1970). Remarks on normalization. In Jacobs, R. and Rosenbaum, P., editors, *Readings in English Transformational Grammar*.

[Cooper, 2007] Cooper, R. (2007). Copredication, dynamic generalized quantification and lexical innovation by coercion. In *Fourth International Workshop on Generative Approaches to the Lexicon*.

[Copestake and Briscoe, 1991] Copestake, A. and Briscoe, T. (1991). Lexical operations in a unification based framework. In *ACL SIGLEX Workshop on Lexical Semantics and Knowledge Representation*.

[Davidson, 1967] Davidson, D. (1967). The logical form of action sentences. In Rescher, N., editor, *The logic of decision and action*. University of Pittsburgh Press.

[de Groote, 2001] de Groote, P. (2001). Towards abstract categorial grammars. In *Association for Computational Linguistics, 39th Annual Meeting and 10th Conference of the European Chapter, Proceedings of the Conference*, pages 148–155.

[De Groote et al., 2009] De Groote, P., Pogodalla, S., and Pollard, C. (2009). On the Syntax-Semantics Interface : From Convergent Grammar to Abstract Categorial Grammar. In Makoto Kanazawa, Hiroakira Ono, and Ruy de Queiroz, editors, *16th Workshop on Logic, Language, Information and Computation Logic, Language, Information and Computation. 16th International Workshop, WoLLIC 2009, Tokyo, Japan, June 21-24, 2009. Proceedings*, volume 5514 of *LNAI/FoLLI*, pages 182–196, Tokyo Japon. Springer. I. : Computing Methodologies/I.2 : ARTIFICIAL INTELLIGENCE/I.2.7 : Natural Language Processing/I.2.7.2 : Language models.

[Dowty, 1989] Dowty, D. R. (1989). On the semantic content of the notion of 'thematic role'. In Chierchia, G., Partee, B. H., and Turner, R., editors, *Properties, Types and Meaning, Volume II : Semantic Issues*. Kluwer Academic Publishers.

[Fillmore, 1965] Fillmore, C. J. (1965). Toward a modern theory of case. Ohio State University Research Foundation.

[Fodor and Lepore, 1998] Fodor, J. A. and Lepore, E. (1998). The emptiness of the lexicon : Reflections on James Pustejovsky's *The Generative Lexicon*. *Liguistic Inquiry*, 29(2).

[Girard, 1971] 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, editor, *Proceedings Second Scandinavian Logic Symp.*, pages 63–92, North Holland, Amsterdam.

[Girard, 1972] Girard, J. Y. (1972). Interprétation fonctionnelle et élimination des coupures de l'arithmétique d'ordre supérieur. Thèse de Doctorat d'État, Université Paris VII.

[Girard et al., 1989] Girard, J.-Y., Taylor, P., and Lafont, Y. (1989). *Proofs and types*. Cambridge University Press, New York, NY, USA.

[Grimshaw, 1990] Grimshaw, J. B. (1990). *Argument Structure*. MIT Press.

[Gruber, 1965] Gruber, J. S. (1965). *Studies in lexical relations*. PhD thesis, MIT.

[Gupta and Aha, 2003] Gupta, K. M. and Aha, D. M. (2003). Nominal Concept Representation in Sublanguage Ontologies. In *Second International Workshop on Generative Approaches to the Lexicon*.

[Gupta and Aha, 2005] Gupta, K. M. and Aha, D. W. (2005). Interpreting Events Using Generative Sublanguage Ontologies. In *Third International Workshop on Generative Approaches to the Lexicon*.

[Hinderer, 2008] Hinderer, S. (2008). *Automatisation de la Construction Sémantique dans TYn*. PhD thesis, Université Henri Poincaré – Nancy 1.

[Huet, 1976] Huet, G. (1976). Résolution d'équations dans des langages d'ordre $1, 2, \ldots, \omega$. Thèse de doctorat d'état, Université Paris VII.

[Jacquey, 2001] Jacquey, E. (2001). *Ambiguïtés lexicales et traitement automatique des langues : modélisation de la polysémie logique et application aux déverbaux d'action ambigus en français*. PhD thesis, Université de Nancy 2.

[Jayez, 2008] Jayez, J. (2008). Quel rôle pour les facettes ? *Langages*, pages 53–68.

[Marcel Cori, 2002] Marcel Cori, J. L. (2002). La constitution du TAL. Étude historique des dénominations et des concepts. *TAL. Traitement Automatique des Langues*, 43(3) :21–55.

[Marlet, 2007] Marlet, R. (2007). When the Generative Lexicon meets Computational Semantics. In *Fourth International Workshop on Generative Approaches to the Lexicon*.

[Mery, 2009] Mery, B. (2009). Compositionality and the Lexicon (Lexique organisé pour la composition sémantique). In *Journées Sémantique et Modélisation*, Paris, France.

[Mery et al., 2006] Mery, B., Amblard, M., Durand, I., and Retoré, C. (2006). A Case Study of the Convergence of Mildly Context-Sensitive Formalisms for Natural Language Syntax : from Minimalist Grammars to Multiple Context-Free Grammars. Rapport de recherche INRIA RR-6042.

[Mery et al., 2007a] Mery, B., Bassac, C., and Retoré, C. (2007a). A montagovian generative lexicon. In *Formal Grammar*.

[Mery et al., 2007b] Mery, B., Bassac, C., and Retoré, C. (2007b). A montague-based model of generative lexical semantics. In Muskens, R., editor, *New Directions in Type Theoretic Grammars*. ESSLLI, Foundation of Logic, Language and Information.

[Montague, 1974] Montague, R. (1974). The proper treatment of quantification in ordinary English. In Thomson, R. H., editor, *Formal Philosophy*, pages 188–221. Yale University Press, New Haven Connecticut.

[Moot, 2007] Moot, R. (2007). Proof nets for display logic. Technical report, INRIA.

[Moravcsik, 1982] Moravcsik, J. M. (1982). How do words get their meanings ? *The Journal of Philosophy*, LXXVIII(1).

[Muskens, 1996] Muskens, R. (1996). Meaning and Partiality. In Cooper, R. and de Rijke, M., editors, *Studies in Logic, Langage and Information*. CSLI.

[Muskens, 2010] Muskens, R. (2010). New Directions in Type-Theoretic Grammars. *Journal of Logic, Language and Information*, 19(2).

[Nirenburg et al., 1995] Nirenburg, S., Raskin, V., and Onyshkevych, B. (1995). Apologiae Ontologiae. In *Memoranda in Computer and Cognitive Science MCCS-95-281*.

[Nordström et al., 1990] Nordström, B., Petersson, K., and Smith, J. (1990). *Programming in Martin-Löf's Type Theory*. Oxford University Press.

[Nunberg, 1993] Nunberg, G. (1993). Transfers of meaning. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, pages 191–192, Morristown, NJ, USA. Association for Computational Linguistics.

[Parsons, 1989] Parsons, T. (1989). The progressive in english : Events, states and processes. *Linguistics and Philosophy*, 12 :213–241. 10.1007/BF00627660.

[Pinkal and Kohlhase, 2000] Pinkal, M. and Kohlhase, M. (2000). Feature Logic for Dotted Types : A Formalism for Complex Word Meanings. In *ACL 2000*.

[Pustejovsky, 1991] Pustejovsky, J. (1991). The generative lexicon. *Computational Linguistics*, 17(4) :409–441.

[Pustejovsky, 1995] Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press.

[Pustejovsky, 1998] Pustejovsky, J. (1998). "Knowledge is Elsewhere" : Natural Language Semantics meets the X-Files. *Linguistic Inquiry*.

[Pustejovsky and Asher, 2000] Pustejovsky, J. and Asher, N. (2000). The Metaphysics of Words in Context. *Objectual attitudes, Linguistics and Philosophy*, 23 :141–183.

[Quine, 1960] Quine, W. V. O. (1960). *Word and Object*. MIT Press.

[Retoré, 2011] Retoré, C. (2011). Properties of the first and second ordre λ-terms and formulas. Personnal communication.

[Saba, 2007] Saba, W. S. (2007). Compositional Semantics Grounded in Commonsense Metaphysics. In *EPIA 2007*.

[Searle, 1979] Searle, J. (1979). *Expression and Meaning*. Cambridge University Press.

[Smith, 2003] Smith, B. (2003). Ontology and information systems. In Floridi, L., editor, *Blackwell Guide to the Philosophy of Computing and Information*, pages 155–166. Blackwell, Oxford.

[Wilks, 2001] Wilks, Y. (2001). The "Fodor"-FODOR Fallacy bites back. In Bouillon, P. and Busa, F., editors, *The Language of Word Meaning*, Studies in Natural Language Processing. Cambridge University Press.