Sémantique et discours - de la modélisation à l'interprétation

Maxime Amblard

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Sémantique et discours de la modélisation à l'interprétation

Mémoire déposé, présenté et soutenu publiquement le 28 novembre 2016

pour l'obtention de l'

Habilitation à Diriger des Recherches
de l'Université de Lorraine

(spécialité informatique)

par

Maxime Amblard

Composition du jury

Rapporteurs :	M. ASHER Nicholas,	Directeur de recherche CNRS, IRIT, Toulouse
	M. BLACHE Philippe,	Directeur de recherche CNRS, LPL, Aix-En-Provence
	Mme TELLIER Isabelle,	Professeure, Université Paris 3 - Sorbonne Nouvelle
Parrain :	M. Philippe de GROOTE,	Directeur de recherche, INRIA, Loria, Nancy
Examinateurs :	Mme Claire GARDENT,	Directrice de recherche CNRS, Loria, Nancy
	M. Jean-Yves MARION,	Professeur, Université de Lorraine

Laboratoire Lorrain de Recherche en Informatique et ses Applications – UMR 7503

Loria
Laboratoire lorrain de recherche en informatique et ses applications

à Hélène,
pour Esther et Bertille

Remerciements

Je tiens à remercier les membres du jury qui ont accepté d'évaluer le document et de participer à la soutenance.

Je tiens à exprimer ma gratitude aux trois rapporteurs d'avoir accepté la tâche de relecture, puis de rédaction d'un rapport. Je remercie très chaleureusement Isabelle Tellier qui a la patience d'évaluer mon travail, de ma thèse au présent document d'HDR. Je remercie chaleureusement Philippe Blache pour ses remarques constructives et l'éclairage donné à mon travail. Une partie importante de mes travaux s'inscrit dans le prolongement des propositions de Nicholas Asher. Sa présence dans mon jury est un honneur pour moi.

Je remercie évidemment et très chaleureusement Philippe de Groote de m'avoir accueilli dans son équipe et d'avoir accepté d'être mon garant d'HDR. Son travail de relecture et les discussions qui s'ensuivirent ont été certes difficiles, mais véritablement passionnantes. J'ai appris énormément avec lui, et comme le montre ce document, je m'inscris pleinement dans la continuité (continuation) de son travail. C'est un vrai honneur de travailler avec lui.

Je remercie bien évidemment Claire Gardent, d'avoir accepté de participer au jury et de l'avoir présidé. Même si j'avais déjà de nombreuses perspectives, Claire m'en a ouvert de nouvelles. Je remercie aussi très chaleureusement Jean-Yves Marion d'avoir accepté de participer au jury. Si ses thématiques de recherche sont plus éloignées des miennes, je crois que je m'inscris dans une certaine filiation épistémologique avec lui.

Je tiens à remercier tous les membres de l'équipe Sémagramme. Nous passons beaucoup de temps ensemble et nos nombreuses conversations professionnelles et personnelles sont nécessaires à mon équilibre. Cette équipe m'a accueilli chaleureusement et offert un cadre stimulant scientifiquement. Un immense merci à Sylvain et Bruno.

Je remercie Céline Simon pour tous les efforts qu'elle accepte de faire pour nous gérer, à la fois dans les moments de stress en amont et ensuite ceux de dépression. J'en profite également pour remercier les différents personnels du laboratoire de l'accueil à la médiation, en passant par l'organisation des colloques.

Je remercie les nombreux collègues du laboratoire qui font de cet endroit un lieu si stimulant, merci à Yannick, Amine, Yves, les nombreux élus que j'ai fréquentés dans les différents conseils.

Un grand merci à Manu et Michel, mes co-auteurs. Il me tarde vraiment de reprendre le travail maintenant que ce document est terminé.

La MSH a également joué un rôle important pour moi ces dernières années. D'abord en accueillant le projet SLAM et puis par extension dans la gestion de la situation (dramatique) qu'elle a traversé. J'y ai rencontré des gens pleinement investis dans leur travail et de grande qualité. Un grand merci à eux, rassuré quand je vois que la structure a su retrouvé un cap.

Bien entendu, la vie scientifique est également faite de tout ce qu'il se passe sur le volet de l'enseignement qui réserve tout à la fois des moments de grande joie et de profonde solitude. Je remercie tous les membres de l'UFR, et en particulier Manu (encore !), Matthieu, Armelle, Laure et Isabelle. Un grand merci aux secrétaires qui travaillent dans l'ombre.

J'ai également beaucoup fréquenté les conseils centraux où je me suis beaucoup amusé et où j'ai beaucoup travaillé. J'y ai rencontré de nombreuses personnes, collègues et personnels administratifs avec qui nous avons passé de longues et ennuyantes, et passionnantes heures. Merci à eux. J'en profite pour très sincèrement et chaleureusement remercier Françoise, nos conversations militantes et scientifiques ont été un soutien important, et m'ont permis de rester ouvert sur l'environnement de travail de nombreux collègues, voire de me rassurer dans les derniers moments.

Un merci tout spécial à Marie C., à la fois pour nos longues discussions sur le fonctionnement de l'Université et pour ses (courageuses) relectures des premières versions.

Je m'excuse évidemment auprès de tous ceux que j'ai oublié de citer.

Un immense merci à Hélène pour sa patience avec moi et tout son amour. Qu'elle sache que c'est réciproque. Merci pour tout son travail de relecture et de correction , cette fois je crois que c'est bon, je ne vais plus rédiger de manuscrit. Je conclus en envoyant mille pensées vers Esther et Bertille qui ont parfois été bousculées par le rédaction de ce manuscrit et la préparation de la soutenance (qu'elles ont écouté plusieurs fois avec attention).

Résumé

Les travaux présentés se situent dans le champ de la linguistique computationnelle. Nous proposons des outils et méthodes informatiques pour le traitement de la langue naturelle. Nos activités de recherche se répartissent selon deux axes :

1. **la modélisation sémantique par des approches formelles et logiques**. Pour cela nous définissons des grammaires respectant le principe de compositionnalité de Frege, s'inscrivant dans la continuité des propositions de Montague, et inspirées par (de Groote 2006) qui propose un calcul sémantique basé sur le λ-calcul simulant la dynamicité.
2. **la confrontation de ces modèles sémantiques et discursifs à des données empiriques** extraites d'usages pathologiques identifiés dans des entretiens entre des patients schizophrènes et des psychologues.

Dans la première partie, nous revenons sur nos travaux en modélisation sémantique. Nous avons été conduits à proposer un formalisme rendant compte de l'interface syntaxe-sémantique pour la théorie générative de Chomsky. Ces grammaires, appelées grammaires minimalistes catégorielles, sont basées sur une extension des grammaires de Lambek, (Lambek 1958), et synchronisent un calcul sémantique au calcul syntaxique par une correspondance entre les types, en s'appuyant sur le λ-calcul. Ce cadre nous a par la suite permis d'interpréter linguistiquement les propriétés de commutativité.

Nous avons ensuite travaillé à la représentation sémantique, ce qui nous a conduit à encadrer deux thèses avec Philippe de Groote. Dans sa thèse, Sai Qian a cherché à modéliser les notions d'événements, de négation et de subordination modale. Une solution pour traiter ces problèmes a été de les envisager comme des problèmes d'accessibilité de variables dans un cadre dynamique. Il a pour cela profondément étendu la notion de contexte de (de Groote 2006).

Nous nous sommes ensuite employés à unifier les traitements dans un unique cadre. Pour cela, Jirka Maršík a, dans sa thèse, proposé un calcul inspiré des propriétés des langages de programmation modernes, notamment les effets algébriques (*effects* et *handlers*). Ce calcul permet de simuler différents ordres d'évaluation, et donc de gérer de manière flexible la notion de contexte. Jirka Maršík a d'une part étudié les propriétés du calcul et prouvé la préservation de types, la confluence et la terminaison, et d'autre part il a montré comment rendre compte de différents phénomènes linguistiques.

Dans la seconde partie, nous nous sommes interrogés sur l'adéquation de ces approches formelles et leur utilisation pour résoudre des problèmes empiriques. La modélisation d'entretiens entre des patients schizophrènes et des psychologues a été le terrain d'étude qui s'est présenté et cela a donné lieu au projet SLAM (Schizophrénie et Langage :

Analyse et Modélisation). Dans ces entretiens, nous avons identifié des échanges dont l'interprétation sémantique ou pragmatique était difficile voire impossible. Le principe a été d'utiliser des formalismes logiques pour la représentation du discours afin d'interroger ces dysfonctionnements.

En étudiant ces entretiens, il nous est apparu pertinent de les analyser sur d'autres niveaux que la sémantique. Nous avons mis en œuvre des outils du traitement automatique des langues sur nos données pour analyser les productions de disfluences, ainsi que la répartition des catégories morpho-syntaxiques. Nous avons ainsi pu identifier que les schizophrènes produisaient plus de disfluences que les interlocuteurs du groupe contrôle.

Finalement, nous avons travaillé à l'utilisation des marqueurs explicites de relations de discours dans des tâches d'extraction d'informations.

La partie finale de ce document revient sur nos perspectives de recherche qui proposent d'unifier les deux axes précédents. Il s'agit de parvenir à réconcilier la vision calculatoire de la modélisation sémantique avec ses applications dans des perspectives des sciences cognitives. Nous souhaitons principalement développer des grammaires sémantiques et la modélisation formelle des dialogues.

Mots clés :
Traitement automatique des langues, logique, langage, grammaires ;
λ-calcul, continuation, modélisation formelle ;
syntaxe, sémantique, discours, pragmatique.

Abstract

Our research is concerned with computational linguistics, proposing computational tools and techniques for natural language processing. Our research activity is spread over two areas :

1. **semantic modeling using formal and logical approaches**. We define grammars that respect Frege's compositionality principle, following the ideas of Montague semantics and inspired by (de Groote 2006), who introduced a theory of dynamics based on the λ-calculus.
2. **the confrontation of these models of semantics and discourse to empirical data** extracted from pathological uses in conversations between schizophrenics and psychologists.

First, we look back on our work in semantic modeling. We proposed a framework for the syntax-semantics interface in the context of Chomsky's generative theory. The grammars, which we call Minimalist Categorical Grammars (MCG), are based on an extension of Lambek grammar, (Lambek 1958), and they coordinate the syntactic and the semantic calculus by a correspondence on types (based on the λ-calculus). This framework has later enabled us to linguistically interpret the commutative properties of the underlying logic.

We then worked on semantic representations, which led us to supervise two PhDs with Philippe de Groote. In his PhD, Sai Qian proposed a model of events, negation and modal subordination. The solution to address these problems has been to consider all of them as phenomena of the accessibility of variables in dynamic semantics. For this, Sai has deeply expanded the notion of context (de Groote 2006).

As a result, we were interested in unifying the treatments in a single setting. In his PhD, Jirka Maršík has defined a calculus inspired by modern programming languages, particularly the use of effects and handlers. The resulting calculus allows us to simulate different orders of evaluation, and thus give flexibility to the context. Jirka has first studied the properties of his calculus and proven subject reduction, confluence and termination. Second, he showed how to use it to account for different linguistic phenomena.

In the second part, we questioned the adequacy of these formal approaches and their use in solving empirical problems. We carried out a field study modeling conversations between schizophrenics and psychologists, which resulted in the SLAM project (Schizophrenia and Language : Analysis and Modeling). In these interviews, we found exchanges whose semantic or pragmatic interpretation was difficult or impossible. The idea was to use logical formalisms for the representation of speech to question these dysfunctions.

When studying these interviews, it seemed appropriate to analyze other levels than

semantics. We implemented automatic processing tools for our data to analyze disfluency production and the distribution of part-of-speech tags. We were able to confirm that schizophrenics produced more disfluencies than interlocutors from the control group.

Finally, we worked on the use of explicit markers of discourse relations in information retrieval.

The final part of the Report highlights our research perspectives that propose to unify the two previous axes. That is, to reconcile computational semantics with its applications in the field of cognitive science. We would primarily develop semantic grammars and formal models of dialogue.

Key words :
Natural Langage Processind, logic, language, grammars ;
λ-calculus, continuation, formal modeling ;
syntax, semantics, discourse, pragmatics.

Table des matières

Table des figures

Liste des tableaux

J'ai un problème de superposition. Si je ne peux pas supperposer deux bouteilles d'huile, il n'y en a pas. Si je n'ai pas la bouteille qui attend que celle qui est en cours ne soit terminée, je suis sans huile. S'il n'y a plus que le sel qui est dans la boîte à sel, sans un kilo de sel dans le placard, je suis sans sel. Ça me rend la vie infernale. Je suis toujours obligée de vérifier deux fois tout. C'est vrai je suis comme ça. C'est un caractère malheureux. C'est un truc arithmétique, c'est 1=0, 1+1=1. C'est ça l'équation, enfin le mode de calcul que je fais dans ma tête.

Marguerite Duras

Introduction

Ce document rassemble les travaux de recherche que nous avons effectués depuis une dizaine d'années. La thématique principale en est la linguistique informatique. Nous nous sommes principalement intéressés à rendre compte de la sémantique de la langue en utilisant la logique. La notion de langue a été définie relativement à celle de langage par Ferdinand de Saussure (de Saussure 1916), qui a eu une influence conséquente sur la science du XX^{e} siècle. Selon lui, le langage est la capacité des humains à exprimer une pensée en s'appuyant sur des signes qui sont organisés en un système spécifique. La langue que nous utilisons en tant qu'humain est riche en ambiguïtés. Même si nous parvenons à les résoudre, ces dernières ont toujours posé de nombreux problèmes dans les théories les modélisant, de la grammaire de Panini aux théories contemporaines.

Nous nous intéressons à l'automatisation des traitements sur des données exprimées dans une langue. Pour y parvenir, il nous faut disposer de représentations qui modélisent les ambiguïtés, afin d'implémenter des programmes qui les reconnaissent automatiquement. On notera que tous les niveaux linguistiques ne posent pas les mêmes natures de difficultés. On parle alors de linguistique informatique ou de traitement automatique des langues (TAL). L'interaction humain-machine directement en langue naturelle est un bon exemple de mise en situation de ces problématiques, car elle mobilise plusieurs niveaux de représentation, bien qu'elle puisse laisser croire que les systèmes sont plus capables que ce qu'ils sont, (Amblard 2013). Cependant, la production de données numériques en langue naturelle est aujourd'hui généralisée, et ces outils trouvent de nombreuses applications pratiques. Nous nous sommes plus particulièrement intéressés à la sémantique de la langue naturelle.

Si nous précisons que notre objet est la langue naturelle [1], c'est qu'il est également possible de définir des langues non naturelles (ou artificielles). Elles nous intéressent aussi car elles permettent de travailler sur des concepts ou des objets spécifiques. Pensons à la langue utilisée en mathématiques qui permet d'exprimer des propriétés difficiles à énoncer en langue naturelle, là où une équation les rend immédiatement intelligibles. En reprenant l'exemple de *loglangue* proposé par (M. Crabbé 2000), « le produit de la somme de deux nombres par leur différence est identique à la différence des carrés de ces nombres », nous préfèrerons écrire que $(n+m)(n-m) = n^2 - m^2$. Un autre exemple, que nous manipulons largement par la suite, est le langage de la logique qui est motivé par une nécessité d'abstraction. Ces langues non naturelles sont construites à partir d'une grammaire qui ne génère pas d'ambiguïtés.

On fait remonter l'étude computationnelle de la sémantique à R. Montague (Mon-

1. Dans la suite, nous utiliserons le terme « langue » qui n'est pas ambiguë en français, contrairement à l'anglais.

tague 1970a ; Montague 1970b) qui a avancé l'hypothèse selon laquelle la sémantique de la langue naturelle suit les mêmes principes que celle des langues artificielles. En cela, il serait possible de rendre compte d'un fragment de l'anglais par le biais de théories et résultats formels, et plus particulièrement des développements de la logique intensionnelle. Pour Montague, contrairement aux grammaires génératives de Chomsky (Noam Chomsky 1957 ; Noam Chomsky 1981 ; Noam Chomsky 1995) à partir desquelles nous avons beaucoup travaillé, la syntaxe est réalisée en surface (sans transformation). Les deux théories s'accordent pour définir le caractère universel de la grammaire. En s'appuyant sur le calcul syntaxique, Montague synchronise l'application de règles de composition. Il est donc en mesure de produire une représentation logique de la langue suivant le principe de Frege (Frege 1892). Il faut attendre (van Benthem 1986 ; van Benthem 1988) pour voir l'introduction d'un calcul effectif, basé sur l'isomorphisme de Curry-Howard (Howard 1980), qui fait la correspondance entre la syntaxe (les grammaires catégorielles) et la sémantique (le λ-calcul (Curry 1934)) à partir des types.

Si cette approche est particulièrement efficace pour produire des représentations sémantiques logiques, il convient de préciser que de nombreuses questions autours de la sémantique demeurent. De manière classique, on peut définir trois types de problématiques pour la sémantique, fortement liées, mais intervenant sur des niveaux différents.

Un premier aspect relève de la sémantique lexicale. Elle s'attache à définir les sens possibles d'un mot, et d'en choisir un en fonction du contexte. De manière schématique, dans l'énoncé « Il commence un livre », l'interprétation finale de « commencer » dépend fortement des caractéristiques de la personne. Si celle-ci est un écrivain, « commencer » signifiera « écrire », si c'est un imprimeur, « éditer », et sinon « lire ».

Un deuxième aspect est concerné par la sémantique compositionnelle qui s'intéresse à comment les items de sens se composent pour former la représentation. L'exemple de la correspondance faite à partir de l'isomorphisme de Curry-Howard relève pleinement de ce niveau. Aujourd'hui, ces questions s'inscrivent dans la définition de l'interface syntaxe-sémantique.

Un troisième aspect rassemble le traitement des phénomènes purement sémantiques comme la résolution des anaphores, des présuppositions ou encore des modalités. Edmond Bach désignait ce niveau et son ouverture vers la pragmatique comme étant la « métaphysique de la sémantique » (Bach 1986 ; Asher 1993).

À la suite des travaux de Montague sur la sémantique, de nombreuses recherches se sont poursuivies tentant d'améliorer la représentation et de faire le lien avec la pragmatique. L'une de ses promoteurs a certainement été B. Partee (Barbara Hall Partee et Hendriks 1997). Dans une autre veine, des recherches se sont inspirées de la théorie des types, comme, sans chercher à les citer de manière exhaustive, (van Benthem 1986 ; Moortgat 1988 ; Morrill 1994 ; Ranta 2004). Un problème classique rencontré par ces approches est d'intégrer une interprétation dynamique de la représentation. L'exemple prototypique se retrouve dans les *donkey sentences*, voir exemple (1).

(1) *Every farmer who owns a donkey, beats it*

La représentation de cet énoncé voudrait qu'un quantificateur existentiel $\exists$ soit utilisé pour l'âne, mais d'une part la variable utilisée par le verbe se retrouve en dehors de la portée de ce quantificateur, et d'autre part l'interprétation de la formule est erronée par rapport à son contenu naturel. La solution est alors d'étendre la portée du quantificateur, et de le remplacer par un quantificateur universel $\forall$.

Des questions s'ajoutent si l'on considère non plus seulement l'énoncé pour lui-même mais dans un contexte global, ce qui devient rapidement nécessaire, par exemple pour résoudre les pronoms personnels anaphoriques (il, elle, ...). Cela constitue l'une des motivations de la définition de plusieurs formalismes comme la *Discourse Representation Theory* (DRT), (Kamp et Reyle 1993). Si cette dernière permet de représenter de nombreux phénomènes à partir de cette proposition, il est difficile de définir un formalisme entièrement calculable, et donc computationnel. Plusieurs formalisations ont été avancées pour rendre compte du caractère dynamique de la langue comme *Dynamic Predicate Logic* (DPL) (Groenendijk et Stokhof 1991), *Predicate Logic with Anaphora* (PLA) (Dekker 1994), le travail de (van Benthem 1995) ou des propositions compositionnelles de la DRT (Muskens 1996).

Un autre système modélisant la dynamicité de la sémantique à partir du λ-calcul a été présenté dans (de Groote 2006) où les termes sémantiques sont étendus pour intégrer le contexte de leur utilisation. Ces principes s'inspirent de la théorie des continuations qui interprète un calcul en fonction de son contexte d'évaluation. Les termes peuvent ainsi agir sur la suite de l'évaluation car ils reçoivent un contexte gauche (ce qui les précède) et un contexte droit (ce qui les suit).

Par ailleurs, si la représentation sémantique est d'une grande utilité pour la compréhension d'un texte, elle ne suffit pas. Il est important, en plus de représenter le contenu informationnel d'un énoncé, d'expliciter comment, et en quoi, il est relié à son contexte. On parle alors de structure rhétorique, le discours ainsi constitué est généralement envisagé au travers de la sémantique, (Asher et Lascarides 2003), plutôt que du point de vue communicationnel (Grosz et Sidner 1986), ou suivant les deux (Mann et Thompson 1988). La *Segmented Discourse Representation Theory* (SDRT) qui est basée sur la DRT, reprend les problèmes de dynamicité et de hiérarchie dans le discours.

Nos activités de recherche sont reprises et présentées les unes en relation avec les autres dans la figure 1. En observant ce schéma, on voit apparaître notre positionnement. Le sujet de notre étude est la langue et plus particulièrement les phénomènes linguistiques qui apparaissent au niveau sémantique. Nous nous intéressons spécifiquement à la représentation de la sémantique par la logique et les outils formels. Il s'agit bien ici de se placer du point de vue de l'informatique pour simuler ces phénomènes linguistiques. Nous cherchons également à interpréter les modèles obtenus par rapport à leur mise en œuvre effective, ce qui relève plus de l'interprétation pragmatique et de la modélisation cognitive.

Dans une première partie de notre projet, nous avons travaillé sur la question de l'interface entre syntaxe et sémantique, et plus particulièrement sur la théorie générative

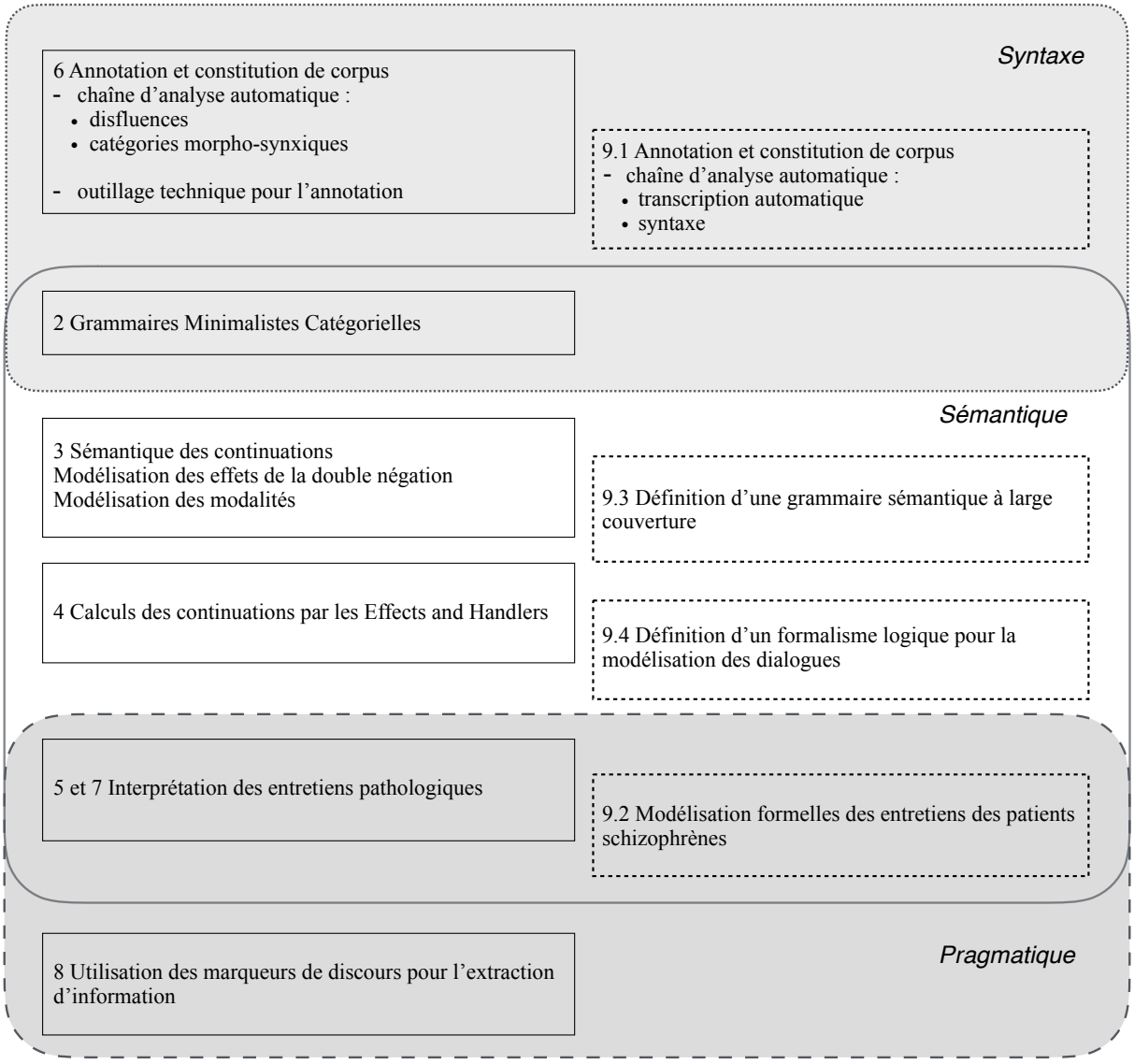


Figure 1 – Schéma général de la structure du document : modélisation par rapport aux trois principaux niveaux linguistiques utilisés, syntaxe - sémantique - pragmatique. Les numéros font références aux chapitres

de Chomsky (Noam Chomsky 1957 ; Noam Chomsky 1981 ; Noam Chomsky 1995). L'une des motivations premières est que cette théorie est largement développée par la communauté linguistique sans être basée sur une formalisation. Le travail a consisté à concilier l'approche montagovienne de la sémantique avec l'approche générativiste de Chomsky comme nous l'avons indiqué précédemment. Le choix fait est substantiellement différent des propositions de Barbara Partee qui se positionne sur cette même question (Barbara Hall Partee et Hendriks 1997 ; Portner et Barbara H. Partee 2002). Pour cela nous avons défini un formalisme dans la tradition logique, basé sur les grammaires catégorielles de Lambek (Lambek 1958), les grammaires minimalistes catégorielles (ou *Minimalist Categorial Grammars* - MCG). Nous avons démontré plusieurs propriétés formelles de la logique mise en œuvre (normalisation faible) et étudié la relation que les MCG entretiennent avec les grammaires minimalistes originelles (inclusion des langages engendrés). Nous avons ensuite défini un calcul sémantique synchronisé avec l'analyse syntaxique, formalisant une interface syntaxe-sémantique.

Cependant, ces premières propositions rencontraient des problèmes dans la gestion du contexte et pour la construction de la représentation du noyau verbal. Nous avons alors tiré parti de l'ensemble des propriétés de la logique (qui possède à la fois des opérateurs commutatifs et non-commutatifs) pour modéliser la notion de phase (Noam Chomsky 1999). Les propositions de (de Groote 2006) ont également été intégrées à l'interface ce qui a permis de proposer un calcul à la fois très simplifié, respectant la compositionnalité et utilisant toutes les propriétés de la logique.

À la suite (et en parallèle) de ces travaux sur la représentation sémantique pour la théorie générative, nous nous sommes intéressés à la modélisation sémantique. En intégrant l'équipe Sémagramme, nous avons débuté un travail sur les *Abstract Categorial Grammars* (ACG), (de Groote 2001), ainsi que sur la théorie des types et le λ-calcul pour la modélisation de la sémantique (de Groote 2006). Une partie de ces travaux a donné lieu à l'encadrement de la thèse de Sai Qian avec Philippe de Groote, où deux extensions ont été définies pour la gestion de l'accessibilité des référents de discours, l'une en présence de négations multiples, l'autre sur la question de la subordination modale (Qian 2014a ; Qian, de Groote et Amblard 2016). Analysons brièvement un exemple caractéristique où l'accessibilité des référents permet de résoudre (ou non) une anaphore pronominale. Soit l'exemple suivant :

(2) (a) Jean a une voiture
    (b) Jean n'a pas de voiture
    (c) Il n'est pas vrai que Jean n'a pas de voiture

(3) Elle est rouge

Dans 3, l'anaphore pronominale « Elle » peut être résolue si l'énoncé précédent est 2(a) ou 2(c), mais pas 2(b). Un important travail a également été produit sur la modélisation de la subordination modale dans ce cadre en modifiant profondément la structure représentant le contexte dans les termes.

Si nous avons pu proposer des modélisations fines de phénomènes linguistiques en conservant un cadre compositionnel, il est apparu qu'il était particulièrement difficile de

définir un cadre qui unifie toutes les propositions. Mais pour construire une grammaire sémantique à large couverture, il nous faut envisager de très nombreux phénomènes. Nous avons encadré une seconde thèse avec Philippe de Groote, celle de Jirka Maršík (Maršík 2016), pour proposer des environnements qui traitent l'intégration de différentes propositions. Il a d'abord cherché à agréger différentes contraintes sur les types des entrées lexicales. Bien que cette solution fonctionne, elle reste trop complexe pour être mise en œuvre effectivement. Une perspective plus fructueuse a été de s'inspirer des propriétés des langages de programmation modernes, en reprenant des effets algébriques (effets de bord (*effects*) et *handler*[2]), (G. Plotkin et Pretnar 2009). L'idée est de disposer d'un environnement qui contrôle l'ordre d'évaluation des calculs. Pour cela, Jirka Maršík a défini un λ-calcul basé sur ces principes (avec une monade[3] simulant les effets de bord et les *handlers*). Il a montré les propriétés fondamentales que sont la préservation de type[4], la confluence[5] et la terminaison[6], qui permettent de définir une forme normale unique. Jirka Maršík a proposé des modélisations pour le traitement de phénomènes sémantiques classiques comme la deixis, la quantification, les implicatures conventionnelles, l'accessibilité des référents de discours pour la résolution des anaphores pronominales, la présupposition, (Maršík et Amblard 2016). Enfin, il a montré comment composer ces différents traitements dans une grammaire sémantique unique (Maršík 2016).

L'ensemble de ces recherches utilise des concepts de l'informatique et de la logique pour la linguistique formelle. Nous participons à une autre recherche basée sur l'utilisation de ces modèles dans des contextes réels. Nous avons intégré et pris la responsabilité scientifique d'un projet interdisciplinaire qui s'intéresse à la modélisation de transcriptions[7] d'entretiens entre des patients schizophrènes et des psychologues. Ces représentations sont questionnées par le prisme de la linguistique, de la philosophie et de l'informatique.

Le projet recoupe plusieurs aspects, mais nous avons plus particulièrement travaillé sur deux volets. Le premier est l'utilisation de formalismes pour la représentation du discours afin d'interpréter les dysfonctionnements qui apparaissent dans ces entretiens. Nous avons utilisé la SDRT (Asher et Lascarides 2003) augmentée d'informations thématiques, ce qui nous a permis de mettre en avant les arguments expliquant pourquoi les structures ne peuvent pas être construites. Dans un second volet, nous nous sommes intéressés à la mise en œuvre des outils du traitement automatique des langues pour analyser quantitativement les productions linguistiques. Nous avons ainsi pu identifier que les schizophrènes présentaient une aptitude à la disfluence légèrement supérieure à celles des autres interlocuteurs. Dans le même temps, ils ont une production normale en ce

2. Une traduction française est « gestionnaire d'événement ».

3. En théorie des catégories, une monade est une construction qui simule le comportement des monoïdes en algèbre. Pour les langages de programmation, une monade est une structure permettant de simuler les effets de bord.

4. La préservation de type assure que l'évaluation d'une expression ne modifie pas son type.

5. La confluence assure qu'en cas d'ambiguïté d'application de règles, pour toute application, il existe une autre règle de réduction amenant au même résultat.

6. La terminaison assure que dans tous les cas, le calcul produit un résultat.

7. La transcription est le processus qui permet de passer d'un enregistrement à une version écrite d'un échange en langue naturelle.

qui concerne les catégories morpho-syntaxiques (*part-of-speech*) ou la diversité lexicale. Nous avons également utilisé ce contexte pour développer l'outillage méthodologique et technique pour porter des campagnes d'annotations manuelles.

Dans la perspective de l'interprétation des représentations sémantiques que nous serions théoriquement capables de produire, nous avons aussi travaillé à l'extraction d'information ou la fouille de texte avec la FCA (*Formal Concept Analysis*) (Ganter, Franzke et Wille 2012), en utilisant des marques explicites de relations de discours comme données d'entrée pour les algorithmes de recherche d'information.

Le dernier point clôt la présentation de notre activité de recherche, mais il reste tout de même, une activité transverse à mentionner relative à l'éthique de la recherche. Il s'agit de considérer les conséquences de la recherche, en particulier du TAL, sur l'organisation de la société et les relations inter-humains, (Amblard, Fort, Musiol et al. 2014). Les techniques et problématiques ne sont pas à juger pour elles-mêmes, mais les finalités doivent nous interroger (Amblard en soumission). C'est dans cette perspective que nous avons participé à l'organisation de journées d'étude sur l'éthique dans le TAL (nationales et internationales), ainsi qu'à la mise en place d'une plateforme de publications sur ce thème [8], qui a par exemple permis de porter une enquête sur les pratiques et la connaissance des scientifiques de l'implication éthique de leurs travaux (Fort et Couillault 2016).

L'ensemble de ces travaux laissent de nombreuses questions ouvertes. Dans le dernier chapitre, nous revenons sur notre projet de recherche qui se découpe en trois axes.

(1) Le premier reprend les propositions actuelles autour du développement de grammaires sémantiques. Il est fréquent de proposer une solution pour le traitement d'un phénomène sémantique particulier, souvent complexe, mais de manière isolée. Et si nous pouvons multiplier les traitements individuels, les rassembler en une unique grammaire implique de gérer les interactions (multiples) entre les différentes solutions, tout en s'assurant de ne pas sur-générer. Une fois que le cadre est posé, il convient de s'intéresser plus profondément au développement du lexique pour augmenter la couverture de la grammaire car une grande partie de l'information utilisée est lexicalisée. Dans notre cas, nous utilisons des ACG que notre problématique nous pousse à étendre, par exemple en s'inspirant des propriétés des langages de programmation (comme Jirka Maršík l'a fait avec les effets de bord et les *handlers*). Une piste en cours de développement est de partir de lexiques existants comme ceux pour les TAG (*Tree Adjoining Grammars*) et de les transformer en lexiques pour les ACG. Plusieurs difficultés apparaissent dans la transformation, notamment autour de la transformation de types qui ouvre plusieurs pistes de recherche.

(2) Un second axe réside dans le prolongement du projet SLAM. L'un des enjeux est de constituer de nouveaux corpus pour affiner les analyses proposées, et d'autre part de les étendre à d'autres pathologies. Il convient de poursuivre sur la question de la modélisation, par exemple en portant de manière plus large des campagnes d'annotations en sémantique et discours pour valider un consensus sur le type de représentation que nous produisons. Pour réaliser ces deux tâches, il faut un important dispositif humain

8. http://www.ethique-et-tal.org/

et une bonne coordination. Par ailleurs, si l'outil développé pour l'analyse automatique, SLAMtk, est aujourd'hui stable, il conviendrait d'augmenter le nombre de phénomènes reconnus et analysés, notamment vers la syntaxe. Une autre perspective est d'utiliser cet outil sur des documents d'autres types du projet SLAM (discours politiques, émission de radio, *etc.*) pour obtenir des analyses automatiques.

(3) Un autre sous-volet concernant SLAM est de poursuivre la modélisation des entretiens avec des patients schizophrènes en s'ouvrant à d'autres pathologies porteuses de troubles de la pensée.

(4) Le dernier volet tend à rassembler les deux précédents. En effet, le premier s'intéresse à la production de grammaires pour la représentation sémantique, et le deuxième met en avant l'interprétation des représentations formelles des dialogues au travers d'une théorie du discours. Il s'agit alors de proposer un outil permettant de produire des représentations pour le dialogue. Une première piste est de s'inspirer de la dynamicité de (de Groote 2006) pour en appliquer les principes au niveau du dialogue. Les interactions étant complexes, la mise en œuvre de ce type de solution ne peut être réalisée qu'à partir d'une grammaire sémantique à large couverture que nous ne possédons actuellement pas (encore).

La structure de ce document suit la logique de cette présentation. Ainsi, le plan est composé de deux parties : l'une sur la modélisation sémantique, et l'autre sur la représentation des discours et l'interprétation de ces représentations.

Avant d'entrer dans la présentation de la sémantique, le chapitre 1 revient sur une présentation historique de l'utilisation de la logique dans la modélisation de la langue. Il s'agit de proposer une contextualisation générale de nos questions de recherche qui tende à expliquer la vision que nous entretenons sur ces questions. Si les travaux fondateurs sont ceux de Montague (Montague 1970a ; Montague 1970b ; Montague 1973b), l'origine épistémologique de ces inspirations permet de comprendre les choix contemporains.

La partie I revient ensuite sur les questions qui nous ont permis de passer de la syntaxe à la sémantique. Nous avons abordé la question de la modélisation sémantique au sens large. Tout d'abord par l'interface syntaxe sémantique dans nos travaux de thèse (Amblard 2007), puis délaissant la relation directe à la syntaxe (ou à un formalisme syntaxique particulier), nous nous sommes intéressés aux représentations sémantiques par la modélisation logique de phénomènes linguistiques complexes et par l'analyse des propriétés formelles des calculs utilisés. Cette partie revient sur nos travaux sur l'interface syntaxe-sémantique dans le chapitre 2 pour présenter les extensions proposées, puis sur deux aspects plus prégnants de nos recherches, l'un sur les travaux de thèse de Sai Qian sur l'utilisation des continuations pour la modélisation sémantique dans le chapitre 3, et ceux de Jirka Maršík définissant un λ-calcul étendu par une monade pour simuler les effets de bord et les *handlers* pour la sémantique des langues dans le chapitre 4.

La partie II rassemble les questions que nous avons traitées autour de la notion de discours. Il s'agit plus particulièrement de donner des cadres applicatifs ou interprétatifs aux outils conceptuels et formels précédemment développés. Pour cela, nous avons principalement travaillé à partir de corpus d'entretiens avec des patients schizophrènes dans

le cadre du projet SLAM, et montré comment les représentations formelles pouvaient prendre sens dans ce contexte. Ce projet a occupé une part importante de notre activité de recherche ces dernières années. Le chapitre 5 présente le contexte, le chapitre 6 la modélisation formelle et le chapitre 7 les questions d'annotations de la ressource développée. Après avoir parcouru les différentes questions soulevées par ce projet, le chapitre 8 revient sur un autre cadre exploré qui est l'utilisation des informations discursives explicites pour une tâche d'extraction d'information. Enfin, le chapitre 9 introduit les grandes lignes de notre projet de recherche.

Chapitre 1

Pour une histoire croisée : linguistique, informatique et logique

Sommaire

Nos recherches s'intéressent à la linguistique informatique. De manière simplifiée, il s'agit de modéliser par des théories formelles et calculables les phénomènes qui apparaissent dans la langue entendue comme la langue utilisée par les humains dans des processus de communication. Nous situons la problématique à la frontière de plusieurs disciplines : l'informatique, la logique, la linguistique et la philosophie du langage. Chacune d'entre elles apporte un éclairage singulier sur un même objet.

Notre problématique se situe plus spécifiquement du côté de la logique et de l'informatique. Nous revenons sur plusieurs éléments épistémologiques qui expliquent comment les questions de modélisation de la langue ont évolué, impliquant que des choix soient réalisés et, en conséquence, pourquoi nous travaillons aujourd'hui dans cette perspective.

Ces questions relèvent du sens des énoncés exprimés dans une langue. Il existe actuellement deux grandes approches de la sémantique qui suivent des méthodes opposées. De manière schématique, on peut identifier les approches basées sur le calcul numérique - comme la sémantique distributionnelle (Harris 1954) - et les approches basées sur la logique - comme la sémantique de Montague (Montague 1974). Il est difficile de les réconcilier dans un même cadre. Nous n'avons pas travaillé sur le volet numérique de la sémantique et nous nous sommes concentrés sur les techniques inspirées de la logique. De fait, notre vision, et donc notre présentation des objets de recherche, se situe plus naturellement sur ces questions.

Dans ce chapitre, nous débutons en reprenant les principaux éléments qui structurent notre vision de la linguistique (formalisée), de la logique et de l'informatique. Comme les influences entre ces disciplines sont nombreuses, nous ouvrons la présentation de chaque discipline par rapport aux autres. Nous nous arrêtons ensuite sur trois étapes clés apparues au XX^{e} siècle que sont : le tournant vers l'interprétation sémantique de la logique avec les travaux de (Tarski 1944), la mise au clair de liens entre description calculatoire et sémantique de la langue, (Montague 1970a), et la volonté de confronter la description syntaxique à la formalisation (Noam Chomsky 1957). Nous terminons par une présentation des développements récents en linguistique formelle et logique qui sont

à la base de nos thématiques.

Nos travaux tendent à définir un calcul rendant compte de la sémantique de la langue, en particulier en utilisant la compositionalité pour construire des formules logiques. La première section de ce chapitre a pour objet de mettre en relation des travaux qui construisent une histoire expliquant l'apparition de cette problématique. Il ne s'agit pas ici de définir une épistémologie générale autour de ces questions, mais de mettre en avant une dynamique scientifique qui conduit à l'utilisation de la logique pour la description sémantique de la langue. Les travaux mentionnés ne sont que des parties restreintes de l'ensemble des travaux de leurs auteurs. Des noms sont mentionnés, mais d'autres scientifiques ont participé à cette construction intellectuelle. Les contributions des auteurs ne sont pas présentées en détail.

1.1 Pour une histoire croisée

1.1.1 Linguistique (et formalisation)

Les premières formalisations de la langue actuellement identifiées remontent à environ 500 ans avant JC. On retrouve plusieurs formes comme les grammaires des langues sémitiques (groupe de langues parlées dès l'Antiquité au Moyen-Orient) ou la grammaire du Sanskrit de Panini. Dans ce dernier exemple, Panini a cherché à décrire systématiquement et méthodiquement la construction des mots. Le Sanskrit utilise l'agglutination qui compose des mots par l'agglomération de plusieurs racines et de marques explicitant les relations entretenues entre ces racines. La description du fonctionnement passe par la définition d'un système basé sur des règles de calcul.

Au IV^{e} ou V^{e} siècle on retrouve une autre description, toujours calculatoire, avec la grammaire de Théodose d'Alexandrie (qui est à l'origine de l'adverbe et d'une description du système de conjugaison). Au VII^{e} et VIII^{e} siècle on voit également l'établissement d'un système de description systématique de la langue pour l'arabe classique. Il est intéressant de noter que la volonté de définir une grammaire de l'arabe est conjointe avec l'apparition de l'idée d'algorithmique du côté des mathématiques. On passe véritablement de l'idée de calculer un résultat à la description de la méthodologie produisant, étape par étape, le résultat.

Sans vouloir écarter des grammairiens et linguistes célèbres qui ont eu une influence importante comme Pascal et d'autres, nous souhaitons conserver le point de vue de la description de systèmes formels. De ce fait, nous ne présentons pas des résultats comme la grammaire de Port-Royal qui fournit une ressource importante, mais qui a justement écartée les descriptions plus formelles du fonctionnement de la langue. Elle a aussi introduit l'idée d'une grammaire universelle, qui sera reprise plus tard par Chomsky, ce qui la rapproche de certaines autres de nos propositions.

Au XIX^{e}, on retrouve le travail (immense) des frères Grimm, qui, en plus de leur travail de passeurs d'histoires pour enfants [1], ont proposé une étude fine des mutations des phonèmes dans les langues indo-européennes. Un exemple relativement simple est la

1. Est-ce que ce sont vraiment des histoires pour les enfants ?

correspondance entre le 'p' et le 'f' entre le français et l'anglais : père-father, pied-foot, *etc.*, (Peyraube 2002).

Il est finalement fréquent de dater la description formelle de la langue à Ferdinand de Saussure (Genève 1857 – Vufflens-le-Château 1913). Il est considéré comme le fondateur de la linguistique moderne et son œuvre ouvre le XX^{e} siècle. Ses travaux ont été publiés après sa mort en 1916 par ses anciens élèves, dans (de Saussure 1916). Ils fondent ce qui reste appelé le structuralisme [2]. Pour Saussure, une langue est « un système dans lequel chacun des éléments (ou signes) n'est définissable que par les relations d'équivalence ou d'opposition qu'il entretient avec les autres, cet ensemble de relations formant la structure ». Son approche est totalement novatrice et s'attache à décrire de manière rigoureuse le fonctionnement de la langue, définissant une structure décomposée (phonologie, morphologie, syntaxe, *etc*). Il définit aussi l'idée d'une analyse diachronique (rapports entre les mots dans le temps) et/ou synchronique (rapports entre les mots à un moment donné) de la langue.

L'influence de Saussure sur la pensée du XX^{e} siècle est très importante. Il introduit entre autres les concepts de signifiant et de signifié. Le signifié désigne le concept ou la représentation mentale d'une chose, alors que le signifiant désigne la représentation mentale de la forme du signe (par exemple l'image acoustique d'un mot). Ces concepts se rejoignent dans la notion de signe et leur étude est la sémiotique (ou sémiologie). Cette branche de la linguistique est proche de la sémantique qui étudie les concepts d'un point de vue linguistique (les signifiés et signifiants).

Le structuralisme introduit une distinction fondamentale entre langage, langue et parole. Par langage, Saussure entend l'aptitude de s'exprimer au moyen de signes. Or cette capacité n'est pas propre aux langages naturels, mais caractérise une volonté de transmettre une information. Dans ce cas, la langue décrit l'ensemble des signes utilisés dans un processus de communication. C'est cette distinction entre ces termes que nous utilisons dans ce document, en employant exclusivement langue pour les langues parlées par des humains. Par ailleurs, Saussure, qui était un spécialiste de la phonétique, fait la différence entre langue et parole. La parole est la mobilisation concrète des signes linguistiques dans un contexte. Il fait ainsi la différence entre la langue concrète (la parole ou le signifiant) et la langue elle-même (le signifié).

1.1.2 Logique (et langue)

La logique possède un langage propre qui permet d'exprimer et de penser sur les objets de la logique. Un exemple de définition d'un tel langage pour la logique est la « loglangue » que l'on retrouve définie et commentée dans (M. Crabbé 2000). Il est difficile de proposer une présentation exhaustive de l'histoire de la logique (et cela ne saurait être l'objet de cette partie). Elle date au moins de l'époque d'Aristote et sa syllogistique qui mettait en exergue le processus d'inférences dans la langue (voir la figure 1.4), mais en restant au niveau syntaxique, comme nous le verrons dans la suite de ce chapitre.

2. Bien qu'il n'ait jamais utilisé le terme de structure lui préférant celui de système, qui se rapproche à nouveau de notre problématique calculatoire.

La période entre Aristote et l'époque moderne a compté de nombreux logiciens (logique du Moyen-Âge, scholastique, *etc.*), mais l'un d'eux a eu une influence immense : Gottfried Wilhelm Leibniz (Leipzig 1646 – Hanovre 1716). On retrouve largement sa pensée philosophique autour de la question de l'individualité et notamment le concept de monade (partie indivisible constituant l'individu). Son domaine de recherche est large et on continue de lui attribuer quantité de concepts fondamentaux de natures très différentes : les termes de fonction et coordonnées, la notation produit, le terme et la notation différentielle ou encore le symbole d'intégrale. [3] (Struik 1969)).

Leibniz introduit la possibilité de définir un langage universel qui serait un calcul algorithmique lui permettant de calculer avec des nombres premiers (malheureusement la complexité de son calcul était gigantesque). La logique de Leibniz est fondée sur deux perspectives. D'une part tout concept est formé d'un très petit nombre de concepts simples qui sont l'alphabet de la pensée humaine. D'autre part, les concepts complexes sont formés selon un principe de composition syntaxique comparable aux règles de l'arithmétique ou de l'algèbre.

Pour Leibniz toute formalisation doit se faire dans l'analogie avec son objet d'étude. Dans (Leibniz 1685) il introduit l'idée que « le seul moyen de rectifier notre raisonnement est de le rendre aussi tangible que celui des mathématiciens, ce qui permet de trouver nos erreurs, et quand il doit y avoir discussion, il devient possible de dire : allons-y, calculons, et voyons qui a raison ».

Deux sicle après, on peut citer le travail méthodique de Giuseppe Peano (Spinetta di Cuneo 1858 – Cavoretto 1932) qui développa une axiomatisation de l'arithmétique. Les notations mathématiques doivent beaucoup à son formulaire de mathématiques (Peano 1889). Il est l'un des premiers à parler de logique mathématique. Peano aura eu une influence importante sur Russell à la fin du XIX^{e}. Peano était également un linguiste italien qui a œuvré pour définir une langue internationale : le latin sans flexion.

Le grand chantier de Leibniz sera complété quelques siècles plus tard par Friedrich Ludwig Gottlob Frege (Wismar 1848 – Bad Kleinen 1925). Frege est considéré avec d'autres comme Russell, Hilbert, Gödel, Cantor, comme père de la logique moderne. Frege se propose de réaliser l'idée de Leibniz d'introduire un langage de la logique. On lui attribue la logique des propositions et le calcul des prédicats. Cette langue formelle de la logique apporte une vision calculatoire. Un de ses grands projets a été de dériver l'arithmétique à partir de la logique (Frege 1879 ; Frege 1994 ; Frege 1892).

Frege introduit les quantificateurs, la théorie de la démonstration et celle de la définition. Il distingue le sens de la dénotation. La dénotation est l'objet auquel on fait référence, alors que le sens est le mode de la dénotation. « 1+1=2 » est une égalité dans laquelle les éléments « 1+1 » et « 2 » dénotent des objets. L'égalité est vérifiée si l'objet « 1+1 » est identique à l'objet « 2 », donc si « 1+1 » et « 2 » ont la même dénotation. Mais les deux formules n'ont pas le même sens. Dans sa logique, les propositions (de la logique des propositions) dénotent des valeurs de vérité (le vrai ou le faux) et sont analysées comme des fonctions.

En parallèle, Georg Ferdinand Ludwig Philipp Cantor (Saint-Pétersbourg 1845 – Halle

3. qui lui valu un procès resté fameux que lui fit Isaac Newton sur la paternité de cette notation.

918) introduit la théorie des ensembles qui est largement reprise par la logique. Il prouve un théorème qui implique l'existence d'une infinité d'infinis. Or cette hiérarchie d'infinis induit des paradoxes sur la cardinalité des ensembles et leurs sous-ensembles.

Bertrand Arthur William Russell (Trellech 1872 – Penrhyndeudraeth 1970), à la suite de Peano comme nous l'avons mentionné, tente d'appliquer la logique aux fondements des mathématiques, ce qui le conduira à partir des travaux de Cantor à exprimer le fameux paradoxe qui porte son nom (Russell 1905 ; Russell 1921) : l'ensemble des ensembles n'appartenant pas à eux-mêmes appartient-il à lui-même ? (Russell 1905).

— Si on répond oui, alors, par définition les membres de cet ensemble n'appartiennent pas à eux-mêmes, il n'appartient pas à lui-même : contradiction.
— Si on répond non, alors il a la propriété requise pour appartenir à lui-même : contradiction à nouveau.

On a donc une contradiction dans les deux cas, ce qui rend l'existence d'un tel ensemble paradoxal. Ce paradoxe est aussi connu avec l'exemple des barbiers (qui ne rasent que ceux qui ne se rasent pas eux-mêmes). L'intérêt de la formulation du paradoxe de Russell est qu'elle est minimale et ne peut donc pas être réparée. Russell a entretenu une correspondance avec Frege sur ces questions car les implications du paradoxe de Russell remettent en cause les définitions de la théorie de Frege.

Rusell pose, sous l'influence de Frege, l'idée que la logique est la combinaison de propositions qui s'analysent en fonction de leurs éléments. Pour résoudre son fameux paradoxe, lorsqu'il travaillait sur les *principia mathématicae* (Whitehead et Russell 1912), il introduit une notion fondamentale (pour nous) : la théorie des types. Cette théorie introduit une hiérarchie de types, puis assigne un type à chaque entité mathématique. Les objets d'un certain type ne peuvent être construits qu'à partir d'objets situés plus bas dans la hiérarchie. Cette propriété empêche que des cercles vicieux (et des paradoxes) surgissent, risquant de casser la théorie.

Bertrand Russell est également connu pour son apport à la philosophie du langage par sa théorie des descriptions définies. En suivant le principe de Frege qui veut que le sens d'une expression complexe soit une fonction du sens de ses parties, il fait l'hypothèse qu'une phrase s'interprète en fonction du sens des éléments qui la composent. Mais que se passe-t-il lorsque l'un des éléments n'existe pas ? L'exemple le plus connu est l'énoncé :

(1.1) Le roi de France est chauve.

Comme il n'y a pas de roi de France, l'interprétation de cet exemple est pour le moins étrange. La solution de Russell est de l'interpréter comme un tout exprimant qu'il y a un élément, que ce dernier est le roi de France, que rien d'autre n'est le roi de France et que cet élément est chauve. Ainsi, s'il n'y a pas de roi la phrase devient fausse, mais non privée de sens.

Après Russell, et toujours dans le prolongement de Frege, apparaît Rudolf Carnap (Ronsdorf 1891 – 1970) un logicien important. Carnap a été le principal étudiant de Frege et a largement œuvré à la reconnaissance de son travail. Une première étape de son travail a été de proposer une construction logique du monde (ou fonder les connaissances du monde sur la logique) (Carnap 1947 ; Carnap 1952). Une partie importante de ses

travaux le conduira à s'intéresser à la modélisation des modalités. Il a fondé une grande partie de sa pensée sur l'analyse logique de la langue.

Dans la perspective de faire le lien entre description logique et description des propriétés du monde, on retrouve le travail de Hans Reichenbach (Hambourg 1891 – Los Angeles 1953) qui est un logicien connu pour ses travaux de modélisation du temps dans la langue (H Reichenbach 1957 ; Hans Reichenbach 1980). Il a fondé avec Carnap une revue de philosophie des sciences, *Erkenntnis*, toujours en activité.

Cette première partie montre que l'histoire s'est faite par transmission entre plusieurs philosophes et logiciens autour de la question de la définition d'un moyen de représentation par un langage spécifique, qui aurait à la fois la concision et l'efficacité pour la pensée qu'ont les mathématiques.

Cette présentation ne prétend pas être exhaustive. Il conviendrait de mettre en avant les travaux de Kazimierz Ajdukiewicz (Tarnopol 1890 - Varsovie 1963) (Ajdukiewicz 1967) un linguiste et logicien qui a travaillé sur l'introduction d'un calcul non-commutatif pour rendre compte de la sémantique, les grammaires AB. Ces grammaires ont été reprises par Yehoshua Bar-Hillel (Vienne 1915 - Jérusalem 1975) (Bar-Hillel 1953) qui a introduit des grammaires logiques pour la syntaxe. Ces constructions ont été largement reprises pour la modélisation de la langue.

Par ailleurs cette présentation ne mentionne pas David Hilbert qui a joué un rôle primordial pour la logique au XX^{e} siècle en introduisant la théorie de la démonstration (Hilbert 1927). Elle ne mentionne pas non plus Kurt Gödel (Brno 1906 - Princeton 1978) qui a également travaillé dans le sillon de Hilbert pour proposer le fameux théorème d'incomplétude qui porte son nom (Gödel 1931), qui affirme que n'importe quel système logique suffisamment puissant pour décrire l'arithmétique des entiers admet des propositions sur les nombres entiers ne pouvant être ni infirmées ni confirmées à partir des axiomes de la théorie. On voit bien là l'impossibilité d'écrire une histoire de la logique en quelques pages. Mais on note que les motivations viennent régulièrement d'une analogie avec le fonctionnement de la langue.

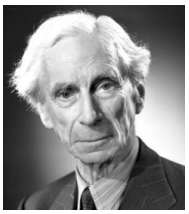
1.1.3 Informatique (et logique et linguistique)

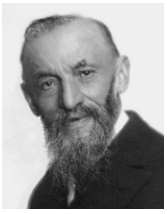
C'est à cette même période, et souvent avec les mêmes influences, qu'on voit émerger les scientifiques définissant la science de l'informatique. Nous avons conclu la partie précédente avec David Hilbert qui est connu pour avoir proposé en 1900 au congrès international des mathématiciens à Paris, 23 problèmes, dont certains ne sont toujours pas résolus (Hilbert 1902). L'un d'entre eux est le problème de la décision : le fait de déterminer de façon mécanique, par un algorithme, si un énoncé se dérive dans un système de déduction sans autre axiome que ceux de l'égalité. Peut-on écrire un algorithme qui décide si un énoncé en logique du premier ordre est valide ? La réponse est négative, mais il faut être en mesure de définir précisément une opération et un calcul.

En fait, il faudra attendre 1936 pour avoir les prémisses d'une réponse à cette question de Hilbert par Alan Mathison Turing (Londres 1912 - Wilmslow 1954). Les travaux de Turing sont aujourd'hui largement connus et diffusés, et ils sont considérés comme fondateurs pour l'informatique. Pour répondre au problème de Hilbert, Turing démontre


(a) Gottfried Leibniz


(b) Gottlob Frege


(c) Bertrand Russel


(d) Giuseppe Peano


(e) Rudolf Carnap


(f) Hans Reichenbach

FIGURE 1.1 – Quelques figures d'une histoire de la logique

qu'il existe des problèmes indécidables (nous ne sommes pas capables de décider s'il est possible de donner une réponse) (Turing 1937 ; Newman et Turing 1942). De manière très schématique, pour y parvenir il introduit un objet abstrait de calcul. Il est défini comme devant simuler un humain calculant, ce qui devra conduire à la construction d'une véritable machine à calculer. Dans le même temps, il introduit l'idée d'une différence entre programmation et programme. Ces principes sont à l'origine de la définition des algorithmes et de la calculabilité, sans lesquels nous ne pourrions avoir d'ordinateur.

Un autre chercheur également à l'origine des ordinateurs est Von Neumann. Il a donné son nom à l'architecture utilisée dans la quasi-totalité des ordinateurs modernes. Il distingue quatre parties : l'unité des traitements, l'unité de séquençage des opérations, la mémoire qui contient à la fois le programme et les données, et un système d'entrée et de sortie (Birkhoff et Von Neumann 1936). Von Neumann est largement influencé par les mathématiques de l'époque, mais a également travaillé sur les aspects d'ingénierie. Van Neumann n'était pas un logicien et c'est là qu'on voit apparaître la nécessaire complémentarité entre spécialistes de la description des calculs formels et ingénieurs capables de les incarner afin de construire une véritable machine de calcul.

Turing est aussi largement cité pour son immense capacité à ouvrir des champs entiers de recherche. Ses travaux, tant en cryptographie qu'en morphogénèse restent fondateurs. Dans notre contexte Turing revêt une autre importance. Il est en effet considéré comme le fondateur de l'intelligence artificielle (IA) avec sa définition du test de Turing (Turing 1950). Ce qui est intéressant est que pour définir l'IA, Turing présuppose le traitement de la langue pour s'attaquer à la question de l'IA. Nous savons aujourd'hui qu'il n'en est

rien, et qu'il convient de redéfinir ce type de test en dehors de la langue naturelle pour ne pas superposer les difficultés.

En 1936, Alonzo Church (Washington 1903 - Hudson 1995) apporte la réponse au problème de Hilbert (Church 1940 ; Church 1951). Church était le directeur de thèse de Turing et c'est lui qui donne le nom de « machine de Turing » à la proposition de Turing. Pour Church, le problème de l'indécidabilité s'incarne dans le λ-calcul, et le corollaire est l'indécidabilité dans le calcul des prédicats. Ce calcul se définit par la notion d'abstraction (l'opérateur λ) et par l'application fonctionnelle :

— λx.M est une fonction sans nom sur le paramètre x et le corps M (abstraction)

— MN est l'appel de la fonction M avec le paramètre N (application)

Le calcul se définit à partir d'une unique règle basée sur la substitution : la β-réduction :

$$(\lambda x.M)N \rightarrow_\beta M[x := N]$$

Cette proposition est à l'origine de la programmation fonctionnelle. Il est intéressant de noter que les réponses différentes à la même question sont à l'origine de paradigmes de programmation très différents et toujours en vigueur (programmation impérative et fonctionnelle).

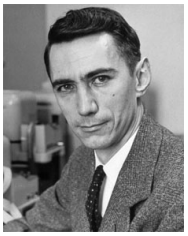
Ces deux propositions sont également à la base de travaux majeurs pour l'informatique qui ont donné la « thèse de Church » (appellation introduite par Stephen Kleene (Hartford 1909 - Madison 1994), (Kleene 1943), puis devenue « thèse de Church-Turing »). Cette proposition est une définition mathématique du concept intuitif de fonction calculable. Turing a montré l'équivalence entre les machines de Turing et le λ-calcul en 1937 (Turing 1937). Ainsi, tout problème de calcul qui peut être résolu par une machine de Turing, peut également l'être par le λ-calcul. Une version plus explicite de sa thèse est que ce qui est explicitement calculable, et en particulier par un système informatique, peut l'être par une machine de Turing. Cette propriété a deux conséquences intéressantes, d'une part tous les langages de programmation sont équivalents et d'autre part que toute théorie suffisante pour capturer les raisonnements mathématiques est incomplète (il existe des énoncés qu'on ne peut ni démontrer, ni réfuter).

Les mathématiques et la logique jouent un rôle important dans la définition de l'informatique théorique. Dans le même temps, la question des langages est intrinsèque à cette science. Elle se décale par rapport aux questions de la linguistique puisqu'elle se concentre sur les langages (non-naturels), mais elle n'exclut pas la question de la langue, comme le montre la définition du test de Turing pour l'IA. Mais plus encore, nous savons que les travaux de Turing furent d'une grande importance pendant la seconde guerre mondiale autour de la cryptographie pour la traduction automatique. Il ne faudrait pas oublier qu'après cette guerre le traitement de la langue va lui aussi être largement influencé par Turing et utiliser des approches analogues aux traitements cryptographiques. Ce type de traduction systématique sera fortement critiqué dans le rapport Alpac en 1966, (Committee 1966). Ce dernier permettra de déplacer le centre d'intérêt de la traduction automatique à la linguistique informatique.

Un autre chercheur généralement considéré comme fondateur de l'informatique est Claude Elwood Shannon (Petoskey 1916 - Medford 2001). Sa proposition principale est


(a) Alonzo Church


(b) Alan Turing


(c) Claude Shannon


(d) John Von Neumann

FIGURE 1.2 – Quelques figures d'une histoire de l'informatique

connue sous le terme de théorie de l'information (Shannon 1949). Il s'agit d'une simplification des systèmes de transmission de l'information et une utilisation des probabilités sur la qualité et la quantité d'information. Ses propositions ont été d'un grand apport pour l'informatique et les mathématiques. Malheureusement, selon lui, ses théories ont été mal reprises en sciences sociales. De fait, la théorie est utilisée pour représenter certaines propriétés des langues. Il reste intéressant de noter que Shannon a débuté sa carrière scientifique en travaillant sur une modélisation électronique de l'algèbre de Boole (et donc de la logique).

Shannon a formulé un modèle de communication suffisamment général pour être mobilisé dans de nombreux domaines. La structure prévoit dès la définition la possibilité de réaliser des traitements mathématiques. Pour lui, tout message se réduit à une description de l'information sous forme binaire. Le théorème qui porte son nom assure qu'on ne perd pas de contenu si le codage a une taille suffisante. Il définit la distinction entre émetteur d'une information et canal par lequel elle est transmise. On définit le codeur pour optimiser la performance source/canal et le décodeur pour optimiser le canal comme transmetteur digital. Shannon a travaillé à développer des calculateurs analogiques qui permettent de connecter émetteur et récepteur par une interface digitale.

La prise en compte de l'information dans la théorie de Shannon n'intègre pas son contenu sémantique, ce qui le différencie de Church et Turing. Il considère uniquement le message par sa forme. Tout comme Turing, il s'est intéressé à l'intelligence artificielle. Il a par exemple travaillé à définir les premiers systèmes capables de jouer aux échecs. Il a du moins montré comment modéliser les parties possibles du jeu d'échecs. On le cite souvent aussi pour sa définition de la machine des machines, composée d'un interrupteur dont l'action est d'éteindre son interrupteur.

1.2 Vers une formalisation de la sémantique des langues

Le cadre dans lequel nous situons nos travaux est largement influencé par les résultats de trois chercheurs du XX^{e} siècle sur lesquels nous revenons maintenant. Ces travaux font

le lien entre langue, langage, logique et interprétation (modèle) et ils concernent à la fois la linguistique, l'informatique et la logique (à des degrés plus ou moins forts). Ces différents éléments sont mis en relation les uns avec les autres dans la figure 1.3.

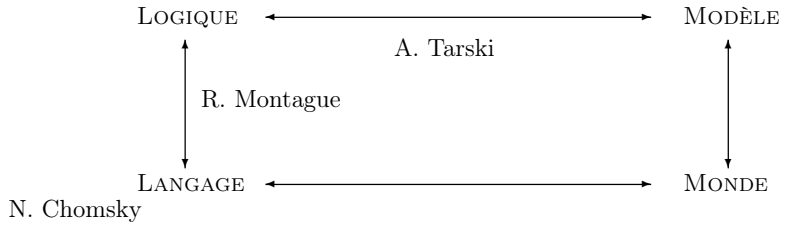


Figure 1.3 – Schéma général de l'interprétation sémantique

Nous revenons d'abord sur le passage à l'interprétation sémantique de la logique proposée par Alfred Tarski. Ensuite, nous introduirons la proposition de l'un de ses étudiants, Richard Montague, qui établit un lien entre sémantique logique et interprétation de la langue. Dans la même période, Noam Chomsky a travaillé à la représentation de la syntaxe par des outils formels.

1.2.1 Le tournant de la sémantique : Alfred Tarski

Autour des années 1930, la vision syntaxique de la logique est délaissée au profit de la sémantique avec le développement de la notion de vérité et de la théorie des modèles, (Tarski 1933). Alfred Tarski (Varsovie 1901 - Berkeley 1983), logicien polonais émigré aux États-Unis en 1939 (A. B. Feferman et S. Feferman 2004), définit le célèbre problème de la satisfaction et introduit la théorie des modèles, (Tarski 1944 ; Tarski 1956). Le principe est qu'une théorie est valide si on peut définir un univers dans lequel elle est vraie. Tarski introduit pour cela la notion de vérité. Ce pas conceptuel est gigantesque car il permet de faire le lien entre la représentation logique et l'interprétation dans une représentation des connaissances (Tarski, Woodger et Corcoran 1956).

La figure 1.4 présente deux exemples de syllogisme dont les premières définitions remontent à Aristote. Les deux respectent le schéma syntaxique de la règle. Le premier exemple est valide alors que l'inférence faite dans le second exemple ne l'est pas (alors qu'il respecte le même schéma). Ce n'est pas que ce second exemple soit faux, mais bien qu'il n'est pas en adéquation avec nos connaissances du monde. La défaillance apparaît dans l'interprétation (le modèle) et non dans l'inférence (la syntaxe de la logique). Il serait possible de définir un modèle d'interprétation dans lequel l'interprétation de ce second exemple soit valide.

Dans la figure 1.3 qui présente le cadre d'interprétation de nos énoncés, nous pouvons situer Tarski comme faisant le lien entre logique et modèle. Nous pouvons déterminer la validité d'une phrase d'une langue dans un modèle à partir d'une formule logique représentant son sens. La qualité de l'interprétation dépendra donc de l'expressivité de la représentation et du modèle censé représenter nos connaissances du monde.

Tous les hommes sont mortels.
Socrate est un homme.

Socrate est mortel.

(a) Syllogisme : exemple de logique dans la langue

Tous les animaux avec des dents sont des pilotes d'avion.
Les poules ont des dents.

Les poules sont des pilotes d'avion.

(b) exemple de la nécessité de la sémantique

FIGURE 1.4 – Exemples de syllogismes

1.2.2 Vers une sémantique formelle de la langue : Richard Montague

À la suite de Tarski, son étudiant Richard Montague (Stockton 1930 - Los Angeles 1971), reprend le lien entre représentation logique et langue. Avant Montague, la sémantique de la langue était principalement l'explication d'ambiguïtés ou de problèmes sur des données souvent subjectives et controversées.

Montague avance l'idée qu'il n'y a pas de différence fondamentale empêchant d'utiliser les outils des langages artificiels pour la langue. Il existe en effet une relation forte car ces langages sont souvent construits par analogie avec les usages de la langue. C'est en particulier le cas pour la logique. Son célèbre article (Montague 1970a) débute par :

« I reject the contention that an important theoretical difference exists between formal and natural languages. »

Je rejette l'affirmation selon laquelle il existe une différence théorique importante entre les langages formels et la langue.

Pour Montague, la langue suit une structure logique. Il s'agit alors pour lui de caractériser les règles de transformation qui permettent de passer de la langue vers le langage de la logique intensionnelle. L'interprétation des formules de cette logique se fait alors dans le cadre de la théorie des modèles. Ainsi chaque élément de sens apporte une contribution à la représentation finale par ses conditions de vérité. Les propriétés syntaxiques servent à construire la forme des énoncés et participent à leur interprétation sémantique grâce à des relations entre des algèbres. C'est une manière d'inclure le principe de compositionnalité de Frege que nous avons énoncé précédemment.

Montague a publié deux autres articles majeurs « Universal grammar » (Montague 1970b) et « The proper treatment of quantification in ordinary English » (Montague 1973b). Pour lui, la notion de grammaire universelle signifiait le développement philosophique et logiquement précis d'une syntaxe, d'une sémantique et d'une pragmatique

englobant à la fois les langages formels et la langue. L'ensemble de ces trois publications introduit un système permettant de construire des représentations sémantiques de la langue par des règles basées sur la logique intensionelle. Il s'est par exemple intéressé à la quantification pour laquelle il introduit la notion de montée de type (*type shifting*). Le type d'un déterminant peut être interprété comme le type d'un quantificateur généralisé.

Montague est mort prématurément et il n'a pas pu traiter de toute la sémantique. Il a cependant introduit de nouveaux traitements pour de nombreux phénomènes sur la quantification, les modifications adverbiales, *etc.* Ses contributions ont été rassemblées dans un important volume (Montague 1974) qui contient également ses travaux faisant le lien avec la pragmatique (Montague 1969 ; Montague 1973a). Sa proposition est d'une grande cohérence formelle et elle permet de prendre en compte l'intensionnalité.

1.2.3 Vers une description formelle de la syntaxe : Noam Chomsky

Dans un mouvement similaire et à une période concomitante, on trouve les travaux linguistiques de Noam Chomsky (Philadelphie 1928) qui mettent en avant l'existence d'une grammaire universelle pour les langues naturelles (Noam Chomsky 1957 ; Noam Chomsky et DiNozzi 1972), dans la veine de Port-Royal. Pour lui la grammaire est abordée comme un organe, au contraire de Montague pour qui la grammaire universelle est un système de relations entre la syntaxe et la sémantique. Le point commun entre ces deux grammaires est certainement qu'elles captent l'idée d'une axiomatisation de la langue naturelle, comme il y a eu axiomatisation du langage des mathématiques.

A contrario de la théorie de Montague, Chomsky suppose que la sémantique ne dépend pas de la syntaxe. Chomsky présuppose l'existence d'un méta-mécanisme fonctionnant comme un tout (l'organe que nous venons de mentionner). Ce mécanisme est alors capable de générer différentes représentations comme une suite de mots (ou leur prononciation) et d'un autre côté une représentation abstraite qui tend du côté de la sémantique (ou de la pragmatique). Les théories de Chomsky sont complexes et interdépendantes (Noam Chomsky 1981 ; Noam Chomsky 1993 ; Noam Chomsky 1999). Pour une présentation détaillée, nous renvoyons le lecteur aux premiers chapitres de (Amblard 2007) qui introduisent les différentes étapes de la théorie chomskyenne.

Ce mouvement pour la constitution d'une grammaire universelle est contemporain de l'essor de l'IA au début de la seconde moitié du XX^{e} siècle qui a influencé Chomsky. Après avoir posé les éléments constitutifs de sa théorie linguistique, Chomsky poursuit par une collaboration forte avec Marcel-Paul Schützenberger (Paris 1920 - Paris 1996) (N. Chomsky et Schützenberger 1963). Les deux scientifiques ont caractérisé les classes de langages engendrées par différents types de grammaires, comme repris dans la figure 1.5. Par ailleurs, ils ont caractérisé les machines reconnaissant chaque type de grammaire. Par exemple, les grammaires régulières sont reconnues par les automates à états finis et les grammaires récursivement énumérables par des machines de Turing.

Ce travail est fondateur de la théorie des langages qui occupe une place prédominante en informatique théorique. Mais il l'est tout autant sur le versant linguistique. En effet, Chomsky s'intéresse à cette question précisément pour caractériser les propriétés des grammaires reconnaissant la langue, et par là même identifier les outils conceptuels

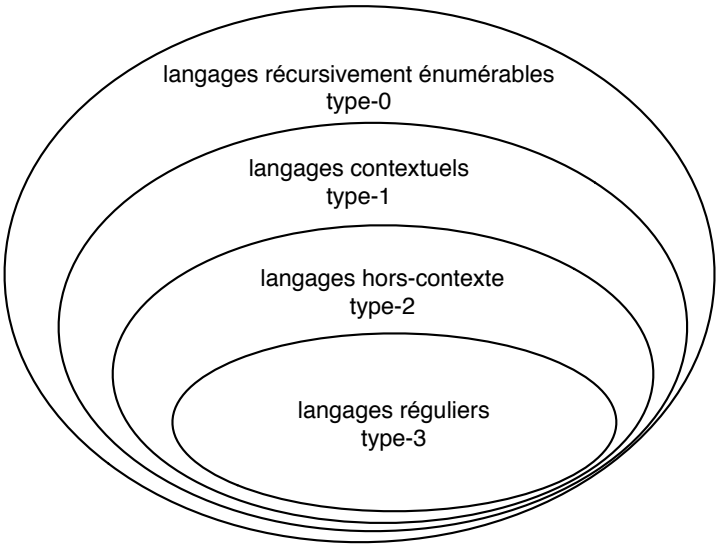


FIGURE 1.5 – Hiérarchie de Chomsky

capables de la reconnaître.

Il apparaît qu'il est malheureusement très difficile de caractériser cette classe. Actuellement on avance qu'elle est *midly context sensitive*, c'est-à-dire qu'elle dépend du contexte sans appartenir entièrement à la classe des langages contextuels. Elle appartient à une sous-classe intermédiaire entre les langages contextuels et hors-contextes. Nous ne disposons donc pas d'un outil abstrait pour la reconnaître explicitement. Un exemple classique utilisé pour montrer que la langue est au delà des langages hors-contextes est les dépendances croisées ($a^n b^m c^n d^m$) que l'on retrouve dans la syntaxe du Suisse-Allemand.

Les influences de Chomsky et de Montague sur la linguistique sont très importantes L'impact sur la sémantique de (Montague 1973b) a été aussi grande que celle de (Noam Chomsky 1957) sur la syntaxe. Bach dans (Bach 1989) résume en disant que la thèse de Chomsky a été que l'anglais pouvait être décrit par un système formel et que la thèse de Montague a été que l'anglais pouvait être décrit par un système formel interprété.

1.3 Inclure la dynamicité à la sémantique

Avant de revenir sur la suite donnée aux propositions de Montague, il convient également de se situer par rapport à l'autre côté du diagramme de la figure 1.3 entre la notion de modèle et de monde. Sur ces questions, de nombreux logiciens ont proposé d'utiliser des extensions logiques particulières pour rendre compte de phénomènes linguistiques. Par exemple Reichenbach, dont il a précédemment été fait mention, ou Johan van Benthem (Rijswijk 1949), membre du groupe Gamut, ont largement travaillé à la modélisation des

notions de temps et aspects qui entrent largement en jeu pour la cohérence linguistique d'un ensemble d'énoncés (H Reichenbach 1957 ; Benthem 2013 ; Stone et Hardt 1999). Un autre exemple est l'utilisation de la logique de Allen, introduite par James Frederick Allen qui est reconnu pour ses contributions en logique temporelle (Allen 1983 ; Allen 1984). Ses travaux sont largement repris dans la définition des théories pour la gestion des événements temps-réels.

Nous revenons à présent sur la formalisation sémantique. On retrouvera dans (Amblard et Pogodalla 2014) une présentation exemplifiée des formalismes introduits dans cette section. Les travaux de Montague ont ouvert des perspectives riches, mais ont été confrontés à des difficultés. Chez Montague, l'analyse syntaxique n'est pas explicitement définie.

Traditionnellement, la notion centrale pour la sémantique logique est que les éléments sont statiques : les prédicats sont interprétés comme des déclarations en fonction d'un modèle. Le contenu sémantique est par ailleurs introduit par le lexique, et une fois introduit dans le calcul, il n'est plus modifié. Un autre point de vue est de ne pas s'intéresser seulement au contenu informationnel, mais de considérer que la manière dont ce contenu apparaît est un objet d'étude. Ces phénomènes relèvent de la sémantique dynamique, comme ce que l'on retrouve dans *Donkey sentences*, (Geurts 2002), repris dans l'exemple 1.2.

(1.2) Every farmer who owns a donkey beats it
Tout fermier qui possède un âne le bât

(1.3) A man walks in the park. He whistles.
Un homme marche dans le parc. Il siffle

En effet, la représentation sémantique de 1.2 doit être la suivante :

$$\forall x.(\mathbf{farmer}(x) \wedge \exists y.(\mathbf{donkey}(y) \wedge \mathbf{own}(x,y))) \rightarrow \mathbf{beat}(x,y) \tag{1.4}$$

Dans la formule 1.4 le y apparaissant dans le prédicat **beat** n'est pas lié par le quantificateur $\exists$ ce qui pose problème. Par ailleurs, l'utilisation du quantificateur $\exists$ ne donne pas une interprétation correcte. La dynamicité permet d'ouvrir à nouveau la portée d'un quantificateur pour le modifier (ici en $\forall$) et d'y introduire une variable :

$$\forall x \forall y.((\mathbf{farmer}(x) \wedge \mathbf{donkey}(y) \wedge \mathbf{own}(x,y)) \rightarrow \mathbf{beat}(x,y)) \tag{1.5}$$

L'exemple 1.3 met en avant un autre problème lié à la construction de structures complexes. L'interprétation de la seconde partie de l'énoncé dépend de la première. Il est donc nécessaire de contextualiser un énoncé pour trouver son interprétation correcte :

$$\exists x.(\mathbf{man}\ x \wedge \mathbf{walk\_in\_the\_park}\ x) \wedge (\mathbf{whistle}\ x) \tag{1.6}$$

Nous sommes à nouveau confrontés à un problème de portée de quantificateurs. Dans 1.6, la variable de **whistle** devrait être liée par le quantificateur correspondant à la variable de l'homme.

Plusieurs solutions ont été introduites pour résoudre ces problèmes. Les travaux les plus influents sont ceux d'Irene Heim, *File Change Semantics* (FCS) (Heim 1982 ; Heim 1983b) et *Context Change Potential* (CCP) (Heim 1983a). Dans ces formalismes l'interprétation est explicitement relative au contexte. Un énoncé n'est plus une valeur de vérité, mais une fonction qui modifie le contexte selon la représentation de son contenu.

À la même période, et sous les mêmes influences, Hans Kamp, étudiant de Montague, introduit la *Discourse Representation Theory* (DRT) (Kamp 1981). Cette théorie déconstruit la notion de portée et la rend plus flexible pour la représentation. Cette solution ne propose pas d'algorithme de construction des représentations, mais est motivée par la définition d'un algorithme de résolution des anaphores pronominales (comme Montague le faisait). La représentation ne porte plus uniquement sur la sémantique mais s'ouvre vers d'autres niveaux de représentation comme le discours (d'où le nom du formalisme), considérant qu'il s'intéresse à la cohérence d'un ensemble d'énoncés.

On trouve d'autres formalismes ayant pour objectif de modéliser le dynamisme comme *Dynamic Predicate Logic* (DPL) (Groenendijk et Stokhof 1990 ; Groenendijk et Stokhof 1991). Dans ce cadre, les formules de la logique du premier ordre sont interprétées comme des ensembles de transitions sur un espace d'états. Pour être plus précis : l'évaluation des formules porte sur des paires d'état d'entrée et d'état de sortie.

En relation avec le problème de la résolution des anaphores pronominales on retrouve la question des présuppositions. Les présuppositions peuvent être vues comme des conditions implicitement validées par un énoncé pour pouvoir comprendre la suite. En cela, les présuppositions relèvent d'un mécanisme dynamique (van Eijck 1995).

Une solution a été proposée par (van Benthem 1991) en faisant la correspondance entre syntaxe et sémantique par les types, contrôlés par l'isomorphisme de Curry-Howard (Howard 1980). Plus précisément, il introduit le lien entre une représentation catégorielle de la syntaxe et un calcul sémantique basé sur le λ-calcul. Cette proposition a véritablement donné une portée calculatoire aux propositions de Montague. Cependant, l'analyse des grammaires catégorielles est relativement limitée. Par exemple, elles ne sont pas capables de reconnaître des dépendances à longues distances. Ainsi l'exemple 1.7, extrait du FrenchTreeBank (A. Abeillé, Clément et Toussenel 2003), contient une extraction médiane d'une subordonnée relative. Ce phénomène n'est pas reconnaissable par une grammaire de Lambek et est difficilement reconnu par les grammaires catégorielles. De fait, il faut envisager d'utiliser d'autres formalismes pour la reconnaissance syntaxique comme nous le proposons dans le chapitre 2.

(1.7) Les investisseurs [$\cdots$] devront échanger leurs devises à un taux que le gouvernement veut situer entre 8 et 10 roubles pour un dollar.

Il s'agit de disposer d'un cadre calculatoire mais également capable de produire des représentations élaborées. On retrouve des propositions pour la modélisation comme (Saurer 1993 ; Van Leusen et Muskens 2002), le traitement d'exemples comme les *Donkey Sentences* dans (Heim 1990) ou (Geurts 2002), ou la prise en compte des notions de temps et aspects (Barbara Hall Partee 1984 ; Kamp, Genabith et Reyle 2011) et de presuppositions (Geurts 1999 ; Beaver 2002). Une autre perspective à la suite de la DRT a été de revenir sur sa capacité de représenter le discours. Par exemple, l'algorithme

sous-jacent pour la résolution d'anaphores donne des résultats erronés. Une solution a été d'enrichir la représentation sémantique par une représentation des relations de discours, ce qui a donné la *Segmented Discourse Representation Theory* (SDRT) (Asher et Lascarides 2003). Le résultat n'est donc plus seulement une représentation logique, mais une représentation structurée de discours à partir de relations coordonnantes ou subordonnantes.

Plusieurs solutions ont été avancées pour définir une sémantique dynamique comme (Zeevat 1989), dans la même veine que (Groenendijk et Stokhof 1990). On retrouve également des tentatives à partir de la modélisation des groupes nominaux (Rooth 1987 ; Barwise 1987). La proposition qui s'inscrit le plus effectivement dans la continuité de (van Benthem 1991) et de la DRT est celle de (Muskens 1996) qui utilise le λ-calcul. Cependant, ces formalismes font face au problème du *destructive assignment*. Ce problème vient de la possibilité de perdre le contenu d'une variable par l'opération d'affectation au cours de l'évaluation d'un programme. C'est le cas de certains langages de programmation : c++, java, *etc.* Une proposition comme (Eijck 1999) permet de disposer d'un calcul sémantique dynamique qui n'a pas ce problème. Une autre solution a été présentée dans *Type Theoretic Dynamic Logic* (TTDL) (de Groote 2006) qui propose une approche purement montagovienne pour le discours en introduisant la prise en compte du contexte dans le λ-calcul. Nous utilisons principalement cette approche qui s'avère très flexible et qui possède une grande capacité de représentation. Nous reviendrons sur la présentation de ce cadre dans le chapitre 3.

Première partie

Modélisation sémantique : de l'interface avec la syntaxe à la composition par les effets

Chapitre 2

Les grammaires minimalistes catégorielles

Sommaire

La relation entre syntaxe et sémantique est au cœur de notre travail de doctorat (Amblard 2007) et des extensions qui ont suivi. L'objectif est de construire une représentation sémantique du contenu des énoncés sous la forme d'une formule logique. Notre problématique se focalise sur la partie compositionnelle de la sémantique, dans le sillage de Montague, où le calcul sémantique est conduit par les relations syntaxiques. Il s'agit d'abord d'identifier une théorie syntaxique.

Un premier élément de réflexion est venu des grammaires de Lambek ou grammaires catégorielles (Lambek 1958 ; Lambek 1961). Ces grammaires sont une formalisation logique de la syntaxe des langues naturelles. En suivant l'idée de (van Benthem 1988), une interface avec la sémantique peut être aisément définie en se basant sur l'isomorphisme de Curry-Howard (Howard 1980) à partir de la notion de types et du λ-calcul. Cependant, la couverture syntaxique de ces grammaires reste très limitée. Il est difficile de rendre compte de dépendances à longue distance, comme dans l'exemple 1.7, qui sont pourtant fréquentes dans la langue.

Une solution a alors été de reprendre la théorie générative de Chomsky, en tant que théorie linguistique, et de travailler à la simuler à partir des grammaires catégorielles. Dans les arguments en faveur de cette théorie, on peut mettre en avant l'importante littérature linguistique, et par ailleurs l'existence d'une première formalisation, connue sous le nom de grammaires minimalistes (*Minimalist Grammars*(MG)), introduite dans (Stabler 1997). Cette formalisation proposait plusieurs éléments pour la sémantique, mais pas un calcul complet.

Nous avons défini les grammaires minimalistes catégorielles (*Minimalist Categorial Grammars* (MCG)), (Amblard 2007), qui reprennent les principes de la théorie générative dans un cadre logique basé sur le calcul de Lambek avec produit, augmenté des connecteurs commutatifs (de Groote 1996). Ces grammaires permettaient de définir une interface avec la sémantique. Récemment, nous avons montré comment modéliser le concept de phase, introduit par Chomsky (Noam Chomsky 1999), en utilisant les propriétés non-commutatives de la logique sous-jacente aux MCG, (Amblard 2011a ; Amblard 2015). Ceci permet d'améliorer la représentation syntaxique et de simplifier le calcul sémantique.

2.1 Définition des grammaires minimalistes

Dans le programme minimaliste (Noam Chomsky 1995), Chomsky propose de rassembler ses théories depuis (Noam Chomsky 1957). La production langagière est vue comme le résultat d'un processus cognitif complexe porté par un substrat commun à tous les hommes, ce qui lui permet de reprendre les principes de grammaires universelles. La modélisation linguistique de ce processus a alors un double objectif : produire une suite de mots qui forme une unité cohérente, et produire une représentation du sens de cette séquence. Ainsi, le résultat rend compte de différents niveaux de représentation linguistique en un seul et même calcul. Dans cette approche, la syntaxe joue un rôle particulier car elle est au centre du processus. Il s'agit de produire une structure complexe qui puisse à la fois se projeter en une représentation de surface et en une représentation sémantique. Les principes généraux de cette théorie sont portés par une idée d'économie maximale, d'où le terme de minimalisme.

Comme nous l'avons introduit dans le chapitre 1, après ses premiers travaux sur la syntaxe, Chomsky s'est intéressé aux aspects formels et aux capacités génératives des grammaires, avant de revenir aux problèmes linguistiques. Il semble pourtant que la théorie générative, qui possède une grande cohérence linguitique, rencontre plusieurs difficultés face à la formalisation. Chomsky n'a d'ailleurs jamais introduit sa propre formalisation de ses théories linguistiques.

La première formalisation des grammaires minimalistes a été introduite dans (Stabler 1997). Le principe est de reprendre les deux opérations de la théorie du minimalisme et de les simuler sur la structure choisie pour la syntaxe par la théorie générative, c'est-à-dire des arbres. Il est relativement aisé de proposer un mécanisme leur associant une forme de surface correspondant à la suite de mots. Par contre, il reste complexe d'intégrer la notion de sémantique dans ces grammaires. On retrouve dans (Heim et Kratzer 1998) ou (Jackendoff 1972) les descriptions des problèmes sémantiques qui émergent. La proposition computationnelle la plus aboutie pour les GM a été définie dans (Kobele 2006). Pour faire le lien entre le minimalisme et les interfaces syntaxe-sémantique basées sur les types, nous avons travaillé à simuler les principes du minimalisme dans un environnement fondé sur des extensions des grammaires de Lambek.

Le minimalisme repose sur deux règles : la fusion (*merge*) et le déplacement (*move*). Il se distingue singulièrement des autres théories en faisant l'hypothèse que les éléments sont mis en relations syntaxiques les uns avec les autres de manière canonique. La dérivation permet de modifier les positions relatives de certaines parties internes. Une illustration classique est celle des questions du type :

(2.1) Quel livre l'enfant lit ?

Dans cet énoncé, l'objet de la phrase a été déplacé de la position canonique pour les objets (à droite du verbe) vers une position devant le sujet.

Les règles de la théorie du minimalisme sont donc :

— *merge* qui permet de rassembler des arbres de dérivation. Par exemple, une première partie du calcul construit la représentation d'un groupe nominal, et l'opération *merge* va permettre de la combiner avec celle d'un verbe.

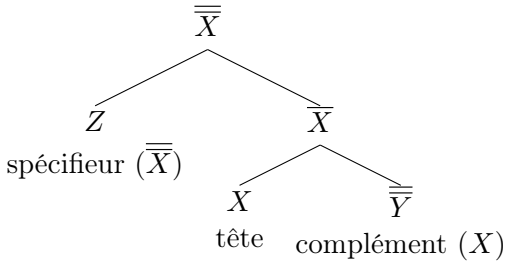


FIGURE 2.1 – Structure générale des syntagmes dans la théorie générative

— *move* qui réalise le déplacement d'un constituant ou d'un syntagme [1] dans une nouvelle position. Ce phénomène est illustré dans l'exemple (2.1). Pour être précis, le déplacement est repris par un élément appelé « trace » qui reste dans la position d'origine. L'example devient donc :

(2.2) [Quel livre]$_i$ l'enfant lit t$_i$?

Nous ne chercherons pas ici à donner tous les détails de la théorie, ni de l'encodage de Stabler ou de celui inspiré des grammaires catégorielles. Le lecteur intéressé trouvera l'ensemble des définitions dans (Amblard 2007). Pour pouvoir présenter les résultats et dresser les perspectives sur ces questions, nous revenons sur quelques éléments de la théorie.

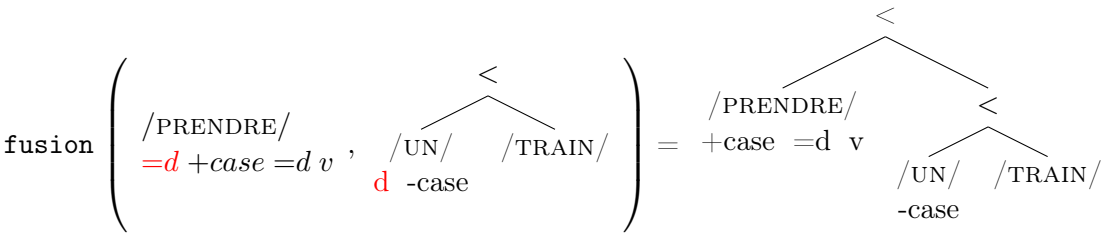
Les arbres manipulés respectent les relations linguistiques internes. Ainsi chaque syntagme est contrôlé par un mot. On parle de gouvernance par un élément lexical qui est nommé la tête (*head*) et qui définit la catégorie principale de la structure. À partir de cette tête, on définit la sous-partie en relation de spécifieur et la sous-partie en relation de complément. La figure 2.1 reprend la structure générale d'un syntagme. Les *compléments* sont considérés comme des propriétés lexicales de la tête. Par exemple, le fait que le verbe *dire* accepte une complétive ne dépend pas de sa catégorie (verbe) mais bien de ses propriétés lexicales. *A contrario*, les spécifieurs sont des propriétés induites par la catégorie de la tête, par exemple le fait que le nom attende un déterminant pour construire un syntagme nominal [2].

Par ailleurs, pour définir correctement le déplacement à l'intérieur des structures, la théorie utilise la notion de projection maximale. Le principe est de déplacer des sous-structures cohérentes (et non n'importe quelle partie). Ainsi, si un critère valide le déplacement, ce dernier est réalisé pour la projection maximale de l'élément. On parle alors de projection maximale d'une tête pour le plus grand syntagme dont elle est la tête, ou

1. Groupe de mots linguistiquement cohérent.

2. Dans la théorie générative, la tête d'un groupe constitué d'un déterminant et d'un nom est le déterminant.

Fusion : le verbe prend un groupe nominal qui sera son objet.



Ce qui permet de déclencher un déplacement :

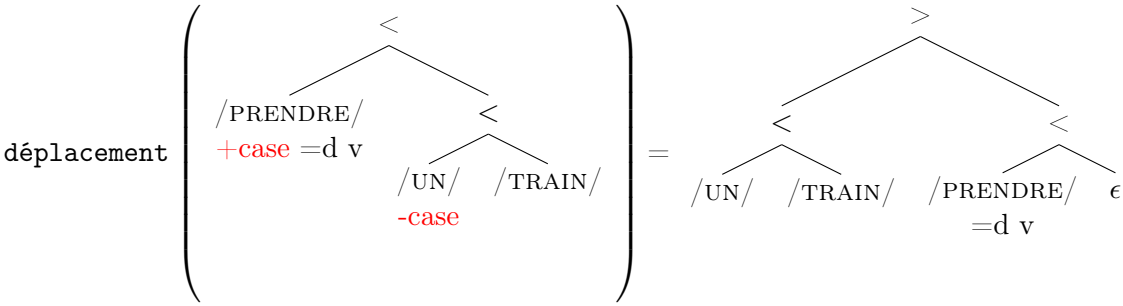


FIGURE 2.2 – Exemple d'application des règles de fusion et de déplacement dans les MG de Stabler

plus structurellement du plus grand arbre d'analyse pour lequel elle reste la tête. L'analyse se fait donc selon trois niveaux : [tête] (niveau 0), [tête [complément]] (niveau 1) et [[spécifieur] tête [complément]] (niveau 2).

Stabler utilise le principe selon lequel les éléments lexicaux portent l'information permettant de contrôler la dérivation. Il s'agit de leur associer des listes de traits (*features*) qui peuvent se combiner entre eux et qui sont consommés par les opérations. Pour conserver l'idée de spécifieur / tête / complément, les arbres contiennent les informations lexicales dans leurs feuilles, et les nœuds montrent la direction vers laquelle se trouve la tête. La figure 2.2 illustre l'application des règles de fusion et de déplacement dans les MG.

Il existe d'autres versions de ces règles qui permettent de préciser l'analyse en séparant la forme phonologique des traits lors des déplacements. Le système est entièrement repris dans (Amblard 2015). À partir de cet exemple on remarque que la partie du calcul correspondant à la production de la suite de mots est simplement la lecture préfixe des feuilles de l'arbre. Par contre, il est plus complexe d'introduire la notion de sémantique sur ces structures.

Dans (Amblard 2007) les principes du programme minimaliste sont posés tant du point de vue linguistique que formel. Dans l'objectif d'introduire le calcul sémantique,

nous avons avons repris les définitions originales pour introduire une caractérisation algébrique qui nous a permis de prouver certaines propriétés du calcul fréquemment admises, en particulier sur la capacité générative de ces grammaires (au travers d'exemples formels).

2.2 Les grammaires minimalistes catégorielles

Dans (Amblard 2007), nous avons proposé une définition des grammaires minimalistes catégorielles (*Minimalist Categorial Grammar* - MCG) en utilisant la logique mixte. Dans nos derniers travaux, nous nous sommes attachés à introduire un concept que Chomsky a ajouté à sa théorie appelé *phase*. Cet ajout implique une modification du formalisme. Afin de simplifier la lecture, cette section reprend les principes généraux des MCG et de l'interface syntaxe-sémantique. La description formelle des règles sera faite en une seule fois dans la section suivante à propos de l'encodage des phases.

Une version des MG dans un cadre logique a été proposée dans (Lecomte et Retoré 2001 ; Lecomte 2005). La notion d'analyse syntaxique (*parsing*) devient la recherche d'une preuve dont la conclusion est la catégorie acceptante, et n'ayant plus d'hypothèse en partie gauche. Ces preuves étant utilisées pour la langue, nous cherchons à leur associer des mots, ce qui est réalisé par un étiquetage.

Les preuves conservent les propriétés des relations syntaxiques de la théorie générative, tout comme les arbres le font dans les MG. Afin de poser le vocabulaire, nous reprenons l'exemple de (Amblard 2015) [3] où l'on cherche à construire une preuve d'un syntagme nominal à partir d'une règle binaire :

$$\frac{\Delta \vdash un : det \qquad \Gamma \vdash livre : nom}{\Delta;\Gamma \vdash un\ livre : SN}\ [\text{règle}]$$

Ici, nous avons une preuve, obtenue à partir de deux prémisses [4] $\Delta \vdash un : det$ et $\Gamma \vdash livre : nom$, qui conduisent à la conclusion $\Delta;\Gamma \vdash un\ livre : SN$ (en dessous du trait horizontal). Chaque élément de cette preuve est un séquent, composé en partie gauche du symbole $\vdash$ d'un multi-ensemble d'hypothèses munies d'un ordre série-parallèle (relations commutatives ou non-commutatives) et en partie droite d'une conclusion, elle-même composée d'une chaîne de caractères et d'une formule. Les relations (comparables ou incomparables) entre les hypothèses en partie gauche des séquents et les conclusions des deux prémisses permettent d'appliquer le schéma de la règle. Ce dernier explicite la formule résultat, c'est-à-dire la manière dont les chaînes de caractères (qui peuvent être vues comme des traits) sont combinées, ainsi que l'impact sur les ensembles d'hypothèses.

La preuve est construite à partir des formules associées aux entrées lexicales des mots de l'énoncé. La théorie générative est interprétée en terme de preuves d'une restriction de la logique mixte (Retoré 2004) (également appelée *Partially Commutative Logic* (PCL)). Cette restriction est, elle, appelée logique minimaliste (*minimalist logic* (ML)). Bien que

3. Cette dérivation ne respecte pas les MCG mais est proposée à titre d'illustration.
4. Une prémisse est une proposition avancée en support à une conclusion.

l'intuition puisse laisser croire que cela complexifie l'analyse, la complexité algorithmique reste polynômiale.

PCL est la logique introduite dans (de Groote 1996 ; Ruet et Fages 1998). Elle a été reprise dans (Retoré 2004) dans une version se concentrant sur les réalisations et que nous utilisons. C'est une superposition du calcul de Lambek (*Intuitionistic Non-Commutative Multiplicative Linear Logic*) et de la logique linéaire intuitionniste multiplicative commutative (*Intuitionistic Commutative Multiplicative Linear Logic*).

Le calcul est fondé sur les connecteurs non-commutatifs du calcul de Lambek ($\odot$, $\backslash$ et $/$) et sur les connecteurs commutatifs multiplicatifs et linéaires ($\otimes$ et $\multimap$). Pour chacun, il existe une règle d'introduction et une règle d'élimination.

L'une des particularités de cette logique est qu'elle est définie pour manipuler simultanément la commutativité et la non-commutativité, en particulier dans les contextes des séquents (en partie gauche). Ainsi, nous manipulons des multi-ensembles partiellement ordonnés d'hypothèses. Deux hypothèses en relation non commutative sont séparées par le symbole ';', et deux hypothèses en relation commutative le sont par le symbole ','. Par exemple $A, B \vdash N$ signifie que pour obtenir un élément de type N, il nous faut deux hypothèses A et B, alors que pour $A; B \vdash N$, il faudra d'abord une hypothèse A avant une hypothèse B. Une règle d'*entropie* définie comme la substitution d'un ordre ';' par un ordre ',' est introduite (notée $\sqsubset$).

Nous faisons l'hypothèse que le calcul des MCG consomme les ressources introduites lexicalement. La ML est uniquement composée des règles d'élimination (privées de $\multimap_e$), de la règle d'axiome et de la règle d'entropie de PCL (voir la figure 2.3). Par exemple, la règle d'élimination de $\backslash$ s'écrit :

$$\frac{\Gamma \vdash A \qquad \Delta \vdash A\backslash C}{\Gamma; \Delta \vdash C}\,[\backslash_e]$$

Cette règle est appliquée à partir de deux séquents, l'un ayant pour conclusion un A, l'autre ayant pour conclusion un $A\backslash C$. Ce dernier doit se trouver sur la droite du premier. La conclusion de la preuve construit le séquent ayant pour conclusion C (où l'on a éliminé le A sur la gauche). Les deux séquents sont réunis par une règle d'un connecteur non commutatif qui ordonne strictement les deux multi-ensembles d'hypothèses (Γ sur la gauche et Δ sur la droite). Ainsi les hypothèses de Δ ne pourront pas commuter devant celles de Γ.

Les règles de tenseurs fonctionnent différemment. Elles combinent une preuve ayant un tenseur en conclusion avec une preuve ayant ces hypothèses dans la partie gauche du séquent (avec un ordre commutatif pour $\odot$ et non commutatif pour $\otimes$). La règle substitue dans la position des hypothèses A et B (ci-dessous en vert) les hypothèses du séquent ayant pour conclusion $A \otimes B$ (ci-dessous en rouge).

$$\frac{\Delta \vdash A \otimes B \qquad \Gamma, (A, B), \Gamma' \vdash C}{\Gamma, \Delta, \Gamma' \vdash C}\,[\otimes_e]$$

$$\frac{\Gamma \vdash A \qquad \Delta \vdash A\backslash C}{\Gamma;\Delta \vdash C}\,[\backslash_e]$$

$$\frac{\Delta \vdash A/C \qquad \Gamma \vdash A}{\Delta;\Gamma \vdash C}\,[/_e]$$

$$\frac{\Delta \vdash A \odot B \qquad \Gamma, A; B, \Gamma' \vdash C}{\Gamma, \Delta, \Gamma' \vdash C}\,[\odot_e]$$

$$\frac{\Delta \vdash A \otimes B \qquad \Gamma, (A, B), \Gamma' \vdash C}{\Gamma, \Delta, \Gamma' \vdash C}\,[\otimes_e]$$

$$\frac{}{A \vdash A}\,[axiome]$$

$$\frac{\Gamma \vdash C}{\Gamma' \vdash C}\,[\sqsubset], \Gamma' \sqsubset \Gamma$$

FIGURE 2.3 – Règles de la logique minimaliste (ML)

Comme nous l'avons introduit, nous ajoutons les règles d'étiquettes sur les preuves afin de disposer d'un calcul sur les mots. Pour cela nous utilisons des étiquetages basés sur des triplets de chaînes de caractères, où chaque position correspond à un élément en position spécifieur / tête / complément de la structure. Un G-étiquetage est une dérivation d'un G-séquent, obtenu par l'application d'une des règles présentées dans la figure 2.4. Le symbole $\bullet$ est utilisé pour la concaténation des chaînes de caractères et $Concat$ est la fonction qui produit la concaténation du contenu d'un triplet. Par exemple, la règle $/_e$ est définie à partir de deux étiquettes r_1 et r_2. Le résultat est le triplet $(\mathrm{r}_{1s},\mathrm{r}_{1t},\mathrm{r}_{1c} \bullet Concat(\mathrm{r}_2))$, c'est-à-dire la première composante est la première composante de r_1, la deuxième composante est la deuxième composante de r_1 et la troisième est le résultat de la concaténation ($\bullet$) de la troisième composante de r_1 et de la concaténation des trois composantes de r_2 (soit $\mathrm{r}_{2s}\bullet\mathrm{r}_{2t}\bullet\mathrm{r}_{2c}$).

La classe de langages reconnues par les MCG contient celle reconnue par les MG (Amblard 2011b). Par ailleurs (Michaelis 2001) a montré que les MG reconnaissent les langages faiblement sensibles au contexte (*midly context sensitive*) - classe supposée des langues naturelles. Enfin (Amblard et Retoré 2014) ont montré la normalisation faible de ce calcul qui assure qu'il est possible de dériver des preuves en forme normale sur lesquelles nous définissons le calcul sémantique. Ce dernier était originellement basé sur le $\lambda\mu$-calculus (Parigot 1992) qui permettait d'utiliser le λ-calcul dans la veine des extensions montagovienne, et de récupérer un calcul proche de ceux inspirés de la théorie des continuations. L'ensemble est présenté en détail dans (Amblard 2007). Nous n'y reve-

Pour s une chaîne de caractères associée à une catégorie A dans le lexique Lex :

$$\frac{\langle s, A\rangle \in Lex}{\vdash_G (\epsilon, s, \epsilon) : A}\,[Lex]$$

Pour x une variable :

$$\frac{x \in V}{x : A \vdash_G (\epsilon, x, \epsilon) : A}\,[axiome]$$

$$\frac{\Gamma \vdash_G r : A}{\Gamma' \vdash_G r : A}\,[\sqsubset], \Gamma' \sqsubset \Gamma$$

$$\frac{\Delta \vdash_G r_2 : B \qquad \Gamma \vdash_G r_1 : B \setminus A}{\langle \Gamma; \Delta\rangle \vdash_G (Concat(r_2) \bullet r_{1s}, r_{1t}, r_{1c}) : A}\,[\setminus_e]$$

$$\frac{\Gamma \vdash_G r_1 : A \,/\, B \qquad \Delta \vdash_G r_2 : B}{\langle \Gamma; \Delta\rangle \vdash_G (r_{1s}, r_{1t}, r_{1c} \bullet Concat(r_2)) : A}\,[/_e]$$

$$\frac{\Gamma \vdash_G r_1 : A \otimes B \qquad \Delta[x : A, y : B] \vdash_G r_2 : C}{\Delta[\Gamma] \vdash_G r_2[Concat(r_1)/x, \epsilon/y] : C}\,[\otimes_e]$$

La règle réalise la substitution de la concaténation de l'étiquette de r_1 en la variable x : $Concat(r_1)/x$.

$$\frac{\Gamma \vdash_G r_1 : A \odot B \qquad \Delta[x : A; y : B] \vdash_G r_2 : C}{\Delta[\Gamma] \vdash_G r_2 \bullet r_1 : C}\,[\odot_e]$$

où $Var(r_1) \cap Var(r_2) = \emptyset$.

FIGURE 2.4 – Ensemble des règles de la ML augmentées des étiquetages

nons pas ici, certaines parties étaient *ad hoc*, et nous les avons modifiées avec la solution présentée dans la section suivante qui introduit le traitement de la notion de phase.

2.3 Encodage des règles : fusion, mouvement et phases

Chomsky introduit après le programme minimaliste la notion de phase (Noam Chomsky 1999). Le processus cognitif de compréhension ne se fait pas exclusivement linéairement, mais fonctionne par étape (par phase), c'est ce qui a motivé l'introduction des phases. Elles constituent des points particuliers dans l'analyse où des propriétés doivent être vérifiées. Dans le cas contraire, il n'est pas nécessaire d'attendre la fin de la production langagière, car elles ne sera pas valide et la compréhension de l'énoncé ne sera pas possible. Réaliser l'analyse par étapes est un avantage pour les systèmes automatiques puisque sans diminuer la complexité des algorithmes, cela permet de limiter la taille des objets sur lesquels les recherches sont réalisées. Chomsky introduit au moins deux phases dans l'analyse d'une phrase standard. La première, vP correspond à l'attribution de tous les θ-roles au verbe, et la seconde CP à un verbe ayant reçu ses informations de forme et de temps.

Plus techniquement, les phases sont constituées d'éléments en position de spécifieur, et d'autres en position de complément. Ceux en position de complément à l'intérieur d'une phase résolue ne sont plus accessibles. Chomsky parle de *Phase Impenetrability Condition* (PIC). Les phases définissent donc des îlots syntaxiques qui restreignent la capacité générative du formalisme. Elles permettent de contrôler les opérations de déplacements et d'éviter les déplacements cycliques infinis. Pour éviter que des éléments soient ainsi fait prisonniers, ils doivent être déplacés en périphérie de la phase avant qu'elle ne soit close.

Dans notre formalisme, les dérivations se construisent autour de la formule associée à l'item lexical du verbe. Cette formule évolue au cours de la dérivation par composition avec d'autres entrées lexicales. Ces éléments qui modifient le verbe permettent justement de définir les différentes parties des phases. Bien que nous ne l'ayons pas énoncé, les MCG définissent un calcul sémantique. Les phases y occupent également un rôle particulier en permettant de gérer les UTAH [5], comme nous le verrons ultérieurement.

Dans les MCG, la fusion est la règle qui combine deux preuves. Pour rendre correctement compte de l'ordre des mots, elle utilise la non-commutativité. Mais dans la même application, la fusion combine également les hypothèses des deux prémisses. D'un point de vue linguistique, les hypothèses fonctionnent au même niveau (à l'intérieur d'un même constituant). Elles doivent pouvoir commuter pour que leur accessibilité soit équivalente, ce que la non-commutativité ne permet pas. La nécessité de mobiliser simultanément la commutativité et la non-commutativité, mais sans leur donner d'interprétation linguistique particulière, a été un argument pour la définition des MCG (Amblard 2007 ; Amblard, Lecomte et Retoré 2010).

Pour réaliser ces opérations dans les MCG, la fusion est une élimination de / ou \

5. *Uniform Theta Assignement Hypothesis* ou assignation de rôles thématiques par le verbe conjugué.

immédiatement suivie de l'application de la règle d'entropie (qui réduit l'ordre non commutatif en commutatif). La fusion se décline en deux versions dépendant du déclencheur gauche ou droit, soit ici de la formule portant le connecteur / ou \. La règle reprenant ces deux applications est notée *mg* (pour *merge*). Pour le calcul sur les chaînes de caractères, la fusion est la concaténation d'une étiquette dans une autre (en fonction de la position gauche/droite).

Déclencheur gauche :

$$\dfrac{\dfrac{\Delta \vdash s : B \qquad \Gamma \vdash (r_s, r_h, r_c) : B \setminus A}{\Delta; \Gamma \vdash (Concat(s) \bullet r_s, r_h, r_c) : A}\,[\setminus_e]}{\Delta, \Gamma \vdash (Concat(s) \bullet r_s, r_h, r_c) : A}\,[\sqsubset]$$

$$\Longrightarrow$$

$$\dfrac{\Delta \vdash s : B \qquad \Gamma \vdash (r_s, r_h, r_c) : B \setminus A}{\Delta, \Gamma \vdash (Concat(s) \bullet r_s, r_h, r_c) : A}\,[mg]$$

Déclencheur droit :

$$\dfrac{\dfrac{\Gamma \vdash (r_s, r_h, r_c) : A \,/\, B \qquad \Delta \vdash s : B}{\Gamma; \Delta \vdash (r_s, r_h, r_c \bullet Concat(s)) : A}\,[/_e]}{\Gamma, \Delta \vdash (r_s, r_h, r_c \bullet Concat(s)) : A}\,[\sqsubset]$$

$$\Longrightarrow$$

$$\dfrac{\Gamma \vdash (r_s, r_h, r_c) : A \,/\, B \qquad \Delta \vdash s : B}{\Gamma, \Delta \vdash (r_s, r_h, r_c \bullet Concat(s)) : A}\,[mg]$$

Par exemple, on suppose que l'on a les séquents $\vdash (\epsilon, un, \epsilon) : k \otimes d \,/\, n$ pour le déterminant et $\vdash (\epsilon, livre, \epsilon) : n$ pour le nom. On les combine par une opération $[mg]$:

$$\dfrac{\vdash (\epsilon, un, \epsilon) : (k \otimes d) \,/\, n \qquad \vdash (\epsilon, livre, \epsilon) : n}{\vdash (\epsilon, un, livre) : k \otimes d}\,[mg]$$

La tête du constituant est le déterminant, comme dans la théorie générative, ainsi que dans la plupart des formalismes fondés sur les grammaires catégorielles. Revenons sur la notion d'hypothèse de nos séquents. La séquence d'hypothèses de la preuve contient exactement la séquence de ressources disponibles pour la suite de la dérivation. Dans ce cas, il s'agit d'un ensemble de ressources et non d'une liste. Cette remarque implique que les hypothèses doivent pouvoir commuter. On applique donc une règle d'entropie dans la définition de la fusion. De plus, cela nous permet de ne pas présupposer un ordre canonique sur la suite d'applications des règles.

Nous supposons l'existence d'items lexicaux, construits sur les hypothèses et la règle d'axiome. Leur contrepartie pour le calcul sur les chaînes de caractères est un triplet

contenant une unique variable. Ainsi pour une catégorie H, la dérivation pourra mobiliser un séquent appelé hypothèse lexicale :

$$H \vdash (\epsilon, x, \epsilon) : H$$

Pour les chaînes de caractères, le déplacement est une substitution. La concaténation de la structure déplacée est substituée à la variable la plus récemment entrée dans la dérivation, l'autre position reçoit la chaîne vide (ϵ) qui correspond dans la réalisation de surface à la trace du déplacement. La règle est notée mv (pour *move*).

$$\frac{\Gamma \vdash r_1 : A \otimes B \qquad \Delta[v : A, u : B] \vdash r_2 : C}{\Delta[\Gamma] \vdash r_2[Concat(r_1)/v, \epsilon/u] : C}\,[mv]$$

En supposant que la dérivation a produit d'un côté notre exemple précédent $\vdash (\epsilon, un, livre) : k \otimes d$ et d'un autre côté $k, d \vdash (v, lit, u) : V$ qui correspond au verbe *lire* ayant reçu deux hypothèses (d et k), il est alors possible d'appliquer mv :

$$\frac{\vdash (\epsilon, un, livre) : k \otimes d \qquad k, d \vdash (v, lit, u) : V}{\vdash (un\ livre, lit, \epsilon) : V}\,[mv]$$

L'analyse n'étant pas terminée, l'ordre des mots peut ne pas être valide. La suite de la dérivation fait évoluer la catégorie du verbe et la chaîne de caractères lui correspondant prend d'autres positions. Les triplets distinguant les positions spécifieur/tête/complément permettent ce type de manipulation.

La règle de phase se décompose en deux parties :

1. le déchargement d'hypothèses en relation non-commutative, noté $[phase]$

$$\frac{\Delta_s, \Delta_h, \Delta_c \vdash (s_s, s_h, s_c) : X \odot Y \qquad \Gamma_s, X; Y, \Gamma_c \vdash (r_s, r_h, r_c) : Z}{\Gamma_s, \Delta_s, \Delta_h \vdash (r_s \bullet s_s, r_h, s_h \bullet s_c \bullet r_c) : Z}\,[phase]$$

2. une partie *transfert* qui réalise tous les déplacements possibles, notée $[phase_t]$. Pour cela la règle $[mv]$ est utilisée.

La règle de phase est donc composée d'une élimination de tenseur non-commutatif, puis de l'application de $[mv]$ (potentiellement plusieurs fois). Une condition est ajoutée sur Δ_c et Γ_c qui doivent être vides après $[phase_t]$ car le contenu interne d'une phase n'est plus accessible en dehors de cette dernière, autrement dit, une dérivation voyant l'une de ses phases ne respectant pas cette condition ne pourra jamais aboutir à une dérivation acceptante (vidée de toutes ses hypothèses). Les phases contrôlent l'analyse syntaxique grâce à cette condition implémentée dans la règle.

(2.3) Un enfant lit un livre

L'analyse de l'exemple (2.3) est reprise intégralement dans la figure 2.6, tant du point de vue de la syntaxe que de la sémantique. Le lexique est constitué des entrées présentées dans la figure 2.5 (où Id dénote l'identité). Nous ne revenons pas en détail sur tous les aspects de cette dérivation, mais nous décrivons le traitement de la première phase (vP).

Une première partie de la dérivation a produit une forme verbale qui peut être composée avec son objet : $d \vdash (\epsilon, lit, u) : (V \odot_< v)$. Le traitement de la phase démarre en combinant cette entrée avec l'entrée lexicale représentant cette phase : $d, k, V; v \vdash (w\ v, \epsilon, \epsilon) : V$. Cette opération est réalisée par une élimination de $\odot$. Nous avons ainsi simulé un point spécifique dans l'analyse qui marque cette phase. Le transfert se produit par la réalisation de tous les déplacements possibles, c'est-à-dire l'introduction de l'objet : $\vdash (\epsilon, un, livre) : k \otimes d$. Ces étapes sont extraites de l'analyse générale et rassemblées dans la dérivation ci-dessous.

$$\dfrac{\vdash (\epsilon, un, livre) : k \otimes d \qquad \dfrac{d \vdash (\epsilon, lit, u) : (V \odot_< v) \qquad d, k, V; v \vdash (w\ v, \epsilon, \epsilon) : V}{d, k, d \vdash (w\ v, lit, u) : V}\ [phase]}{d \vdash (w\ un\ livre, lit, \epsilon) : V}\ [phase_t]$$

L'utilisation de la notion de phase permet de définir une interface sémantique simplifiée. En effet, un des arguments pour leur introduction est associé à la notion de θ-role (rôle thématique). L'une des difficultés de la première interface, était justement de répartir ces informations entre différentes entrées lexicales du verbe (donc différentes variables) et de les unifier par le calcul. La contrepartie sémantique des phases permet d'introduire les prédicats représentant les assignations de rôle, et la réalisation de la phase par une élimination de tenseur, permet d'accéder explicitement aux variables à unifier. Cet argument paraît technique, mais l'unification des variables engagées dans les prédicats des θ-roles dans la première interface, était un problème ouvert.

Par ailleurs, nous reprenons également le calcul dynamique introduit dans (de Groote 2006) et décrit dans le chapitre 3. Pour cela, nous utilisons le λ-calcul avec des types étendus aux contextes. L'interprétation du verbe transitif n'est plus $\lambda xy.verbe(y, x)$, comme dans la théorie montagovienne, mais $\lambda xye\phi.verbe(y, x) \wedge \phi e$. Ici e est la variable représentant le contexte gauche dans lequel le terme est utilisé (ce à partir de quoi il est interprété) et ϕ le contexte droit (ce vers quoi il poursuivra son interprétation). Le contexte peut être vu comme une liste de variables introduites et accessibles pour le reste de l'interprétation, comme dans (de Groote 2006). L'opérateur ' : :' est utilisé comme constructeur de liste.

L'interface syntaxe-sémantique est ainsi très réduite : dans le calcul sémantique, l'application fonctionnelle est associée à la fusion et au déplacement ; la phase utilise aussi l'application fonctionnelle ainsi qu'un terme spécifique qui décharge les variables des deux termes en jeu dans la phase avant de les recombiner. Ce terme dépend du type de la phase (ou dit autrement du nombre de variables dans le terme du verbe) :

— deux arguments : $\lambda T_1 T_2 O S r.S(\lambda x.O(\lambda ye\phi.T_1(\lambda A.A)(\lambda B.B)rxye(\lambda e'.T_2 rye'\phi)))$

— un argument : $\lambda T_1 T_2 S r.S(\lambda xe\phi.T_1(\lambda A.A)rxe(\lambda e'.T_2 rxe'\phi))$

La différence entre ces termes est mise en avant avec la couleur gris. Ces termes peuvent paraître complexes, mais ils ne font que redistribuer les variables sur leurs arguments, ce qui permet d'unifier les différentes variables (contextes, réification, *etc.*) tout en contrôlant l'adéquation des types.

C'est le seul niveau de complexité de cette interface contrairement aux premières versions qui contenaient plusieurs éléments calculatoires *ad hoc*. Seuls les termes lexicalisés

un	$\vdash (\epsilon, un, \epsilon) : k \otimes d \,/\, n$	$\lambda PQe\phi.\exists x.Px(x :: e)(\lambda e'.Qxe'\phi)$
enfant	$\vdash (\epsilon, enfant, \epsilon) : n$	$\lambda ze\phi.child(z) \wedge \phi e$
livre	$\vdash (\epsilon, livre, \epsilon) : n$	$\lambda ze\phi.book(z) \wedge \phi e$
lire	$\vdash (\epsilon, lit, \epsilon) : (V \odot_< v) \,/\, d$	$\lambda OSr.S\lambda x.O\lambda ye\phi.read(r) \wedge \phi e$
mode	$V; v \vdash (\epsilon, \epsilon, \epsilon) : k \setminus d \setminus V$	$\lambda rze\phi.patient(r, z) \wedge \phi e$
infl	$\vdash (\epsilon, -, \epsilon) : k \setminus (c \odot t) \,/_< V$	Id
comp	$c; t \vdash (\epsilon, \epsilon, \epsilon) : c$	$\lambda rze\phi.agent(r, z) \wedge \phi e$

FIGURE 2.5 – Lexique syntaxico-sémantique d'un fragment d'une MCG

participent à la formule finale avec des prédicats. Donc l'interprétation sémantique des hypothèses lexicales est l'identité.

Les termes associés au déterminant et au nom restent traditionnels (augmentés de la notion de continuation), ceux associés aux items lexicaux portant les phases (mode et comp) apportent l'information sur les rôles thématiques correspondant au cas syntaxique qu'ils introduisent dans la dérivation.

Sans revenir en détail sur l'explication de la dérivation qui reconnaît notre exemple, nous revenons sur la contre-partie sémantique de notre phase exemple. Nous renvoyons pour l'analyse générale à (Amblard 2015), tout comme pour un exemple où la phase permet de bloquer une analyse erronée. La fusion et le déplacement sont la simple application fonctionnelle, la dérivation peut construire le terme sémantique associé à la forme verbale : $\lambda OSr.S(\lambda x.O(\lambda ye\phi.read(r) \wedge \phi e))$. La phase en cours de traitement traite un verbe devant être combiné avec un objet et un sujet (soit deux variables). Sa contrepartie sémantique est l'utilisation du terme précédent :

$$\lambda T_1 T_2 OSr.S(\lambda x.O(\lambda ye\phi.T_1(\lambda A.A)(\lambda B.B)rxye(\lambda e'.T_2rye'\phi)))$$

Il est appliqué au verbe, puis appliqué au terme de la phase, soit $\lambda rze\phi.patient(r, z) \wedge \phi e$. En utilisant plusieurs β-réductions nous obtenons le terme de la phase qui est ensuite combiné dans la partie de transfert. Cette dernière est composée de déplacements qui sont eux-mêmes des applications fonctionnelles, ici avec le terme de l'objet : $\lambda Qe\phi.\exists x.book(x) \wedge Qx(x :: e)\phi$.

$$\dfrac{\dfrac{\vdash (\epsilon, un, livre) : k \otimes d \quad \lambda Qe\phi.\exists x.book(x) \wedge Qx(x :: e)\phi \qquad \dfrac{d \vdash (\epsilon, lit, u) : (V \odot_< v) \quad \lambda OSr.S(\lambda x.O(\lambda ye\phi.read(r) \wedge \phi e)) \qquad d, k, V; v \vdash (w\ v, \epsilon, \epsilon) : V \quad \lambda rze\phi.patient(r, z) \wedge \phi e}{\lambda T_1 T_2 OSr.S(\lambda x.O(\lambda ye\phi.T_1(\lambda A.A)(\lambda B.B)rxye(\lambda e'.T_2rye'\phi))) \quad d, k, d \vdash (w\ v, lit, u) : V \quad \lambda OSr.S(\lambda x.O(\lambda ye\phi.read(r) \wedge patient(r, y) \wedge \phi e))}\,[phase]}{d \vdash (w\ un\ livre, lit, \epsilon) : V \quad \lambda Sr.S(\lambda xe\phi.\exists z.book(z) \wedge read(r) \wedge patient(r, z) \wedge \phi(z :: e))}\,[phase_t]$$

On voit dans cet exemple que la gestion des θ-roles n'est pas lexicalisée dans l'item du verbe. Cette stratégie suit la position davidsonienne[6], (Davidson 1967), que nous

6. Nous reviendrons sur cette question dans le chapitre 3.

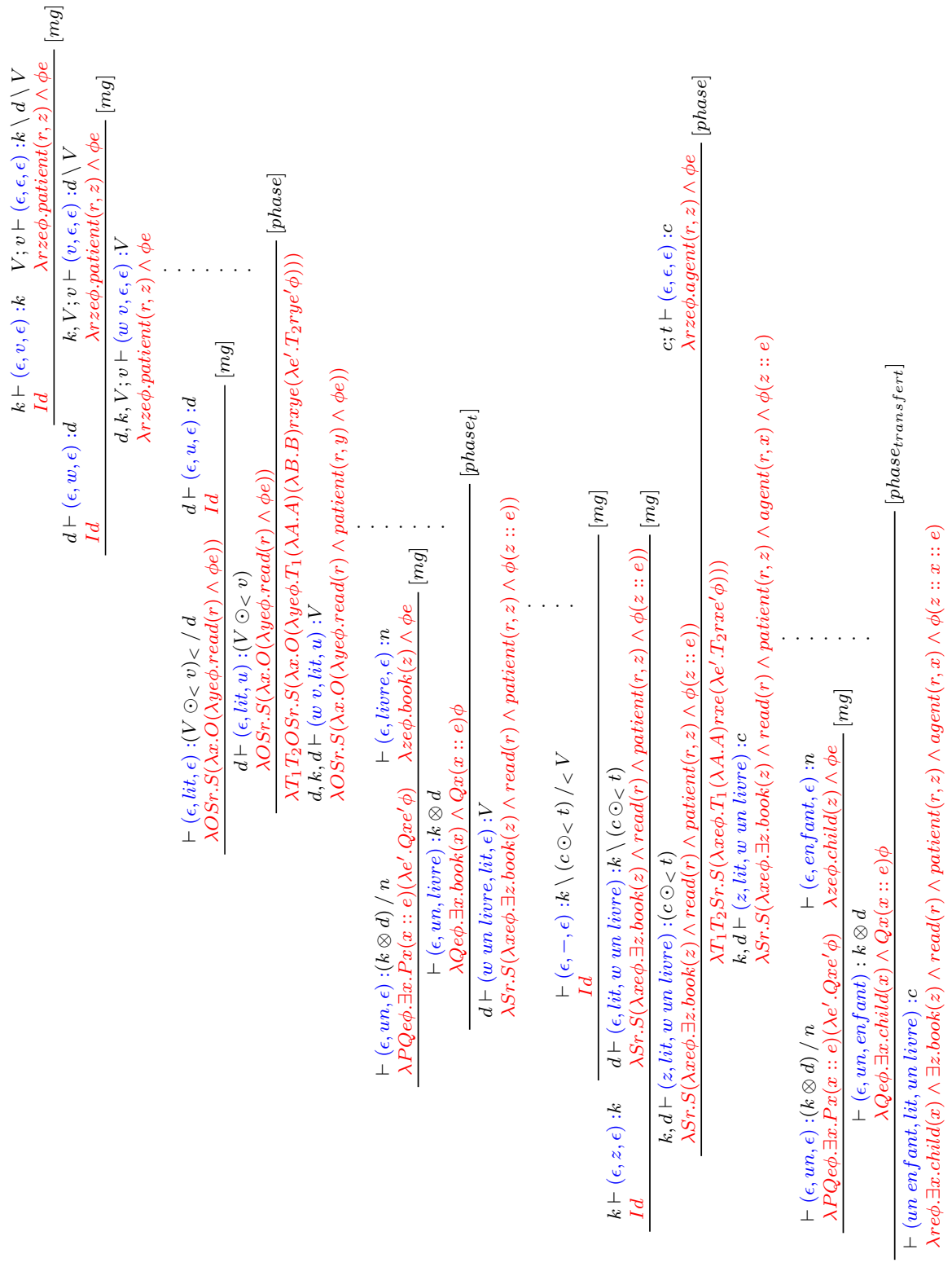


FIGURE 2.6 – Dérivation syntaxique (en noir), sur les chaînes de caractères (en bleu) et sémantique (en rouge) de l'exemple 2.3

avons par ailleurs adoptée pour le verbe. La difficulté est alors d'unifier les variables d'événements, et c'est le rôle du terme de la phase. Ce qu'on ne voit pas dans cette présentation, c'est la simplification du traitement induite par la phase. Elle est simulée au travers d'hypothèses dans la dérivation qui permettent de rassembler différentes variables. La version sans les phases nécessitait plusieurs règles *ad hoc* pour traiter correctement l'unification, (Amblard 2007).

Il convient de noter que le premier calcul sémantique proposé était très complexe. Il s'agissait de passer outre le problème classique des *donkey sentences* en ouvrant à nouveau la portée d'un quantificateur. Pour y parvenir, nous avions intégré une notion de portée explicite comme celle utilisée dans la DRT. Afin d'y accéder, nous utilisions le $\lambda\mu$-calcul de Parigot (Parigot 1992). Ces propriétés permettaient d'obtenir le résultat escompté, mais introduisaient la non-confluence du calcul sémantique. Sans considérer que c'était un problème, cela permettait d'obtenir plusieurs représentations sémantiques en cas d'ambiguïté dans la phrase.

La proposition de (de Groote 2006) est bien plus élégante en ce qu'elle résout le problème de réouverture de la portée (ou de dynanimicité de la quantification) en introduisant directement dans le λ-calcul la notion de continuation.

Sur les aspects de modélisation par les grammaires minimalistes catégorielles, nous avons également travaillé à montrer que la classe de langages reconnue par les MCG contient celle reconnue par les MG Amblard (2011b). La normalisation de PCL proposée dans (Amblard et Retoré 2014) assure qu'il est possible de produire une dérivation acceptante ayant une forme normalisée.

Chapitre 3

Sémantique dynamique

Sommaire

Après nos travaux sur l'interface syntaxe-sémantique pour la théorie générative, nous nous sommes concentrés sur les aspects de modélisation de la sémantique. En plus de la question du transfert des informations structurelles identifiées au niveau de la syntaxe, se pose celle du type de représentation que nous souhaitons produire en sémantique. Un exemple classique est la prise en compte des modifieurs adverbiaux pour lesquels il existe plusieurs solutions, par exemple l'utilisation de la notion d'évènements, (Davidson 2001), sur lesquels nous revenons dans la section 3.2. Nous nous restreignons à la production de formules logiques construites à partir du λ-calcul (ou d'extensions de ce calcul).

Une autre question est de passer d'une description de la sémantique à une représentation plus riche en contenus, prenant en compte par exemple la dynamicité de la langue. Pour illustrer ce dernier point, nous utilisons un court exemple, repris de (Amblard et Pogodalla 2014). La représentation du discours composé des phrases 3.1a et 3.2a produit une représentation telle que celle présentée en 3.3.

(3.1) a A man entered.
b $\exists x.\mathbf{man}(x) \wedge \mathbf{entered}(x)$

(3.2) a He smiled.
b $\exists x.\mathbf{smiled}(x)$

(3.3) A man entered. He smiled.
$\exists x.\mathbf{man}(x) \wedge \mathbf{entered}(x) \wedge \mathbf{smiled}(x)$

Chaque phrase introduit ses quantificateurs comme dans 3.1, et il n'est alors plus possible d'ouvrir leur portée pour y ajouter un prédicat. Or ce phénomène d'ouverture se produit pour obtenir 3.3. De plus, il faut identifier que le pronom *He* fait référence à l'homme introduit dans 3.1 et unifier les variables correspondantes. La dénotation dépend fortement du contexte, ce qui implique de disposer de calculs dynamiques pour la sémantique.

Nous avons choisi de travailler à partir des grammaires catégorielles abstraites (ou *Abstract Categorial Grammars* - ACG), (de Groote 2001). L'utilisation des types permet de synchroniser l'analyse de plusieurs niveaux. De manière traditionnelle, la réalisation de surface est entendue comme l'énoncé lui-même, et la réalisation profonde comme son

analyse syntaxique ou sémantique. Les ACG sont particulièrement adaptées pour passer d'une analyse syntaxique à une représentation sémantique et elles sont basées sur le λ-calcul.

Dans la continuité de la tradition montagovienne (Montague 1970a), (de Groote 2006) a proposé d'intégrer le traitement de phénomènes dynamiques dans le λ-calcul. Il s'inscrit dans la suite des propositions de (Barker 2004) qui voit la sémantique des langues comme des phénomènes de continuation. Pour cela, des contextes sont introduits dans les termes. Cette proposition ouvre une voie pour la modélisation sémantique en ce qu'elle augmente l'expressivité du contenu informationel des formules. Par exemple, il devient possible de simuler en partie la DRT (Kamp et Reyle 1993) et la SDRT (Asher et Lascarides 1998). De nombreuses questions persistent quant à l'utilisation de la modélisation de cette notion de contexte pour la sémantique de la langue.

Nous avons donc travaillé à la modélisation de phénomènes spécifiques en utilisant ces contextes, en particulier dans le travail de Sai Qian, que ce soit dans son master que nous avons encadré, (Qian 2009), et dans sa thèse sous notre direction avec Philippe de Groote (Qian 2014a). La section suivante revient sur les définitions originelles du calcul. Nous présentons ensuite le traitement de phénomènes demandant une adaptation de la structure de contexte (le pluriel, les événements et la négation). Enfin, Sai Qian s'est intéressé à la modélisation de la subordination modale qui a demandé un travail plus conséquent que nous aborderons dans la section 3.3.

3.1 Contextes dans le λ-calcul

Pour étudier la sémantique, nous avons choisi d'utiliser la proposition de (de Groote 2006) qui résout le problème majeur que nous avions rencontré pendant notre thèse. Il s'agit de rendre compte de phénomènes dynamiques de manière compositionnelle en s'appuyant sur la théorie des continuations. On retrouve dans (Amblard et Pogodalla 2014) une présentation exemplifiée des arguments qui ont conduit à passer d'une sémantique statique à des représentations dynamiques complexes.

La notion de continuation a aussi été proposée pour formaliser les flux de calculs dans les langages de programmation (Strachey et Wadsworth 1974). Le principe est de disposer de la suite d'un calcul et de pouvoir l'évaluer relativement à son utilisation. Une continuation est une représentation abstraite d'un état de l'exécution d'un programme. Ces approches présentent l'intérêt de permettre la transmission des résultats après l'évaluation, plutôt que de devoir arrêter une évaluation. C'est par exemple un mécanisme utile pour l'encodage des exceptions dans les langages fonctionnels. Une manière de simuler les continuations est d'ajouter un argument qui représente la suite du calcul aux fonctions.

Les continuations ont été reprises pour la sémantique de la langue pour divers phénomènes linguistiques (Barker 2002 ; Barker 2004 ; Shan 2004 ; de Groote 2006). Dans l'exemple 3.4, l'interprétation de l'objet est très différente en fonction de sa nature (« Marie » ou « tout le monde »). On peut définir la dénotation de « tout le monde » comme la manipulation du contexte. Dans ce qui suit, nous présentons le cadre proposé par (de

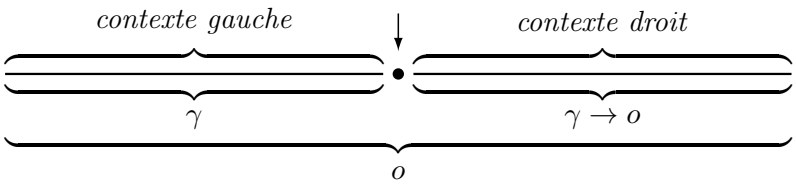


FIGURE 3.1 – Représentation des contextes typés pour un énoncé

Groote 2006), que nous appelons *Type Theoric Dynamic Logic* (TTDL).

(3.4) Jean aime Marie/tout_le_monde

En ajoutant la notion de contexte aux représentations basées sur le λ-calcul, TTDL permet de calculer des phénomènes dynamiques. Les arguments suivants sont identiques à ceux qui ont motivé la DRT, TTDL conservant la compositionnalité du calcul. La dénotation d'un énoncé devient dépendante de son contexte d'apparition, tant de ce qui le précède (le contexte gauche) que de ce qui le suit (le contexte droit).

Ces principes entraînent une modification profonde des types associés aux phrases. Nous utilisons les notations de (Church 1940) pour noter les types : o le type des valeurs de vérité et ι le type des individus. Nous ajoutons γ le type associé aux contextes. La figure 3.1 reprend la présentation faite dans (de Groote 2006). Le point représente un énoncé. L'interprétation globale reste une valeur de vérité o. Nous venons d'argumenter pour la considérer en interaction avec son contexte gauche auquel nous assignons le type γ. Le contexte droit permet donc de passer d'un contexte gauche à une valeur de vérité, soit le type $\gamma \to o$. Grâce à tous ces éléments, nous définissons le type de la dénotation d'un énoncé comme étant : $\gamma \to (\gamma \to o) \to o$. La dénotation d'une phrase est maintenant le point de jonction entre un contexte gauche et un contexte droit. En notant s la catégorie d'une phrase, on obtient :

$$[\![s]\!]_{TTDL} = \gamma \to (\gamma \to o) \to o \tag{3.5}$$

Ces modifications sont reportées au niveau du discours, ce qui donne pour d, la catégorie des discours :

$$[\![d]\!]_{TTDL} = \gamma \to (\gamma \to o) \to o \tag{3.6}$$

Posons que ce nouveau type correspond à celui des propositions dynamiques.

On remarquera qu'il suffit de passer la continuation triviale comme argument pour récupérer l'interprétation sans contexte. Il nous faut donc un mécanisme pour construire le discours à partir des phrases. Nous utilisons pour cela un opérateur qui combine les propositions dynamiques :

$$[\![D.S]\!] = \lambda e\phi.[\![D]\!]e(\lambda e'.[\![S]\!]e'\phi) \tag{3.7}$$

Dans une première approximation, nous considérons que cette notion de contexte est pertinente pour la résolution des anaphores pronominales, tâche à l'origine de la définition de la DRT. Sans chercher à définir un algorithme qui résout explicitement ces dernières, nous pouvons limiter la liste des référents accessibles en fonction du contexte qui est alors constitué d'une liste d'entités introduites dans le discours et accessibles.

Dans (de Groote 2006), le contexte est modélisé à partir de l'opérateur ' : :' qui est un constructeur de liste. Son type est donc $\iota \to \gamma \to \gamma$. Par ailleurs, afin de compléter le système, nous faisons l'hypothèse de l'existence d'un opérateur de sélection `sel` de type $\gamma \to \iota$ qui extrait une entité d'une liste (un individu d'un contexte).

Pour obtenir l'interprétation de l'exemple 3.3, les contextes sont représentés par des variables. On appelle traditionnellement e le contexte gauche et ϕ le contexte droit. Ce dernier est une fonction qui, à partir d'un contexte de type γ, produit une proposition de type o. Le passage d'un contexte gauche à un contexte droit est dénoté par l'application de ϕ à e. Le terme représentant l'exemple 3.1a avant d'être combiné avec ses contextes devient :

$$\lambda e\, \phi.\exists x.(\mathbf{man}(x)) \wedge \mathbf{entered}(x) \wedge (\phi(x :: e))$$

Le terme est très proche, mais on voit que le contexte gauche est augmenté de la variable x, avant d'être passé au contexte droit (en rouge). Le nouveau terme de 3.2b est quant à lui :

$$\lambda e\, \phi.(\mathbf{smiled}(\mathtt{sel}\, e)) \wedge (\phi\, e)$$

Il s'agit simplement de combiner les contextes et d'ajouter un prédicat. De la même manière que nous pouvons rassembler l'interprétation d'une phrase avec un discours pour produire un discours, il est également possible de combiner deux phrases entre elles, avant de les intégrer dans un discours. L'opérateur $\overline{\circ}$ réalise alors le même calcul que dans 3.7.

$$\begin{aligned} [\![S_1.S_2]\!] &= [\![S_2]\!] \overline{\circ} [\![S_1]\!] \\ &= \lambda e\, k.[\![S_1]\!]\, e\, (\lambda e'.[\![S_2]\!]\, e'\, k) \end{aligned} \tag{3.8}$$

Le calcul complet de notre exemple est présenté dans la figure 3.2. Le résultat contient ici les mêmes prédicats sauf pour le prédicat **smiled** qui a comme argument `sel` $(x :: e)$. Le sélecteur est une fonction capable d'extraire une entité dans un ensemble d'entités accessibles. L'ensemble des entités accessibles est construit à partir du contexte gauche par la variable e, et contient également la variable liée x qui est justement la variable recherchée. Par ailleurs, la dernière partie de la formule montre que le contexte gauche e, augmenté de la variable x est combiné avec le contexte droit ϕ. Grâce aux continuations, la portée du quantificateur a été ouverte et la représentation de la phrase suivante a pu y être intégrée.

Cette approche s'inscrit dans la continuité de la sémantique de Montague. Les types des syntagmes qui dépendaient du type des phrases continuent d'en dépendre, bien que le type de la phrase soit modifié en $\Omega = \gamma \to (\gamma \to t) \to t$:

Montague		*Continuation*	
$[\![S]\!]$	$= o$	$[\![S]\!]$	$= \Omega$
$[\![NP]\!]$	$= (\iota \to [\![S]\!]) \to [\![S]\!]$	$[\![NP]\!]$	$= (\iota \to [\![S]\!]) \to [\![S]\!]$
$[\![N]\!]$	$= \iota \to [\![S]\!]$	$[\![N]\!]$	$= \iota \to [\![S]\!]$

$$
\begin{aligned}
[\![AME.HS]\!] &= [\![AME]\!] \bar{\circ} [\![HS]\!] \\
&= \lambda e\, \phi.[\![[\![AME]\!]\, e\, (\lambda e'.[\![[\![HS]\!]]\!]) \\
&= \lambda e\, \phi.[\![[\![AME]\!]\, e\, (\lambda e'.(\lambda e\, \phi.(\mathbf{smiled}(\mathtt{sel}\, e)) \wedge (\phi\, e))\, e'\, \phi) \\
&= \lambda e\, \phi.[\![[\![AME]\!]\, e\, (\lambda e'.(\mathbf{smiled}(\mathtt{sel}\, e')) \wedge (\phi\, e')) \\
&= \lambda e\, \phi.(\lambda e\, \phi.\exists x.(\mathbf{man}(x)) \wedge (\mathbf{entered}(x)) \wedge (\phi(x :: e))) \\
&\qquad\qquad e\, (\lambda e'.\mathbf{smiled}(\mathtt{sel}\, e') \wedge (\phi\, e')) \\
&= \lambda e\, \phi.\exists x.(\mathbf{man}(x)) \wedge (\mathbf{entered}(x)) \\
&\qquad\qquad \wedge (\lambda e'.(\mathbf{smiled}(\mathtt{sel}\, e')) \wedge (\phi\, e'))(x :: e)) \\
&= \lambda e\, \phi.\exists x.(\mathbf{man}(x)) \wedge (\mathbf{entered}(x)) \\
&\qquad\qquad \wedge ((\mathbf{smiled}(\mathtt{sel}\, (x :: e))) \wedge (\phi\, (x :: e)))
\end{aligned}
$$

FIGURE 3.2 – Représentation sémantique par β-réduction de deux énoncés successifs (on note $[\![AME]\!]$ l'interprétation de *A man entered* et $[\![HS]\!]$ celle de *He smiled*)

Il est également possible de dériver une version dynamique des connecteurs de la logique statique. Trois connecteurs seulement sont nécessaires, les autres pouvant être déduits à partir des lois de de Morgan : la conjonction, la négation et la quantification existentielle.

$$
\begin{aligned}
P \bar{\wedge} Q &= \lambda e\, k.P\, e\, (\lambda e'.Q\, e'\, k) \\
\bar{\neg} P &= \lambda e\, k.(\neg P\, e(\lambda e'.\top)) \wedge (k\, e) \\
\bar{\exists} x.P &= \lambda e\, k.\exists x.P\, x\, (x :: e)\, k
\end{aligned}
$$

Enfin, il faut transcrire les propositions simples comme $\mathbf{man}(x)$ dans une version dynamique qui prend en compte la gestion des contextes $\lambda e \phi.(\mathbf{man}(x)) \wedge (\phi\, e)$, nous pouvons déduire le nouveau lexique dynamique pour réaliser la dérivation de l'exemple 3.3, comme présenté dans la figure 3.3. Pour plus d'explications sur la construction automatique des lexiques dynamiques, nous renvoyons à la lecture de (de Groote 2010 ; Lebedeva 2012 ; Qian 2014b). On remarquera que le pronom n'est pas simplement dérivé de sa version statique, sans quoi nous ne pourrions pas résoudre le problème. Enfin, en observant la traduction de $\bar{\neg} P$ on note que $\neg P$ est construit à partir de la continuation triviale, ce qui signifie que $\neg P$ est entièrement évalué sans ce contexte. Dans ce cas, la continuation notée ϕ n'apparaît pas dans la portée de la négation. De plus, le contexte est alimenté par le même contexte que P impliquant que tout référent de discours introduit dans P n'est pas transmis à ϕ [1].

1. Ceci correspond à la contrainte d'accessibilité telle qu'elle est exprimée dans la DRT.

$$
\begin{aligned}
[\![man]\!] &= \overline{\lambda x.\mathbf{man}(x)} \\
&= \lambda x e\, \phi.(\mathbf{man}(x)) \wedge (\phi\, e) \\
[\![a]\!] &= \overline{\lambda P\, Q.\exists x.(P\, x) \wedge (Q\, x)} \\
&= \lambda P\, Q e\, \phi.\exists x.(P\, x\, (x :: e)(\lambda e'.Q\, e'\, \phi) \\
[\![entered]\!] &= \overline{\lambda s.s(\lambda x.\mathbf{entered}(x))} \\
&= \lambda s e\, \phi.s(\lambda x.(\mathbf{entered}(x)) \wedge (\phi\, e)) \\
[\![smiled]\!] &= \overline{\lambda s.s(\lambda x.\mathbf{smiled}(x))} \\
&= \lambda s e\, \phi.s(\lambda x.(\mathbf{smiled}(x)) \wedge (\phi\, e)) \\
[\![he]\!] &= \lambda P e\, \phi.P\, (\mathtt{sel}\, e)\, e\, \phi
\end{aligned}
$$

FIGURE 3.3 – Exemple de transformation d'un lexique statique en une version dynamique

On retrouve les bases des continuations sémantiques pour le discours et les anaphores dans (de Groote 2006). Une construction sur ces bases, tenant compte des présuppositions - comme le font (de Groote et Lebedeva 2010 ; Lebedeva 2012) - a été proposée dans (Martin et Pollard 2010 ; Martin et Pollard 2012).

Cette courte présentation reprend les grandes lignes de l'encodage de la dynamicité et des continuations dans le λ-calcul pour la langue. Cette proposition donne des solutions pour certains problèmes mis en avant pour la sémantique de Montague. Mais plus encore, et bien que nous n'ayons pas insisté sur ce point, cette proposition s'inscrit dans la continuité de la DRT de Hans Kamp, (Kamp et Reyle 1993), tout en conservant la compositionnalité et sans *destructive assignement*. Par ailleurs, comme nous l'avons sous-entendu, l'une des questions suivantes est d'utiliser pleinement cette sémantique pour exprimer des phénomènes sémantiques.

3.2 Modélisation sémantique et structure du contexte

Dans la droite ligne de la théorie des continuations précédente, nous nous sommes intéressés à mobiliser le contexte pour proposer des modélisations linguistiques plus élaborées, comme une modélisation de l'accessibilité des référents de discours pour le pluriel (Qian et Amblard 2012). Pour cela, nous avons proposé une méthodologie qui construit les ensembles d'ensembles (*powersets*) des référents introduits dans la continuation. Ainsi tout sous-groupe d'éléments du discours peut être utilisé pour une référence anaphorique. La solution permet de résoudre les anaphores des exemples suivants :

(3.9) *John was in Paris. Jean was in Rome. Mary was in Barcelona.*

(3.10) *They would come back to work after the vacation.*

(3.11) *They avoided the bad weather in France/Italy/Spain.*

Dans ces exemples le *They* fait référence à des groupes d'individus très différents en fonction du contexte. Nous avons donc classé le problème d'accessibilité des référents de discours entre *summation* et *abstraction* en fonction des différents antécédents. Nous avons proposé des solutions pour chaque type en considérant que les référents individuels et les groupes partageaient le même type. Cette hypothèse apporte une distinction entre les notions d'ensembles traditionnels en mathématiques ou en logique et les objets que nous manipulons, il est alors possible que tous les sous-groupes servent d'antécédents pour la résolution des anaphores. Nous admettons de fait que la taille du contexte croît de façon exponentielle avec le nombre de référents accessibles. Une stratégie pourrait être de mobiliser un apprentissage de la limite de la taille de l'ensemble des référents à maintenir accessibles à partir d'exemples réels.

On notera que cette proposition est assez technique pour modéliser ce seul phénomène. Au delà de l'intérêt formel, cette proposition permet d'appréhender d'autres problèmes sous un autre angle. C'est le cas notamment des quantifications vagues du type « un peu » ou « la plupart ».

Nous avons également travaillé à intégrer les événements dans ces contextes (Qian et Amblard 2011) ce qui correspond à un problème tout à fait classique en modélisation de la sémantique logique. La gestion des modificateurs implique une refonte profonde des représentations.

Une solution à cette question a été apportée par (Davidson 1965 ; Davidson 1967) avec l'idée de réifier les formules. Le principe de la réification est de supposer que les entités peuvent être représentées par des objets. Des concepts abstraits vont ainsi s'incarner dans des objets concrets, ici, la formule que nous sommes en train de construire et une variable. Davidson nomme ces variables des événements. La représentation sémantique revient ainsi à construire des éléments qui participent à la description des événements. Ainsi, l'exemple 3.12 n'est pas représenté par la formule $\exists x.plazza(x) \wedge kiss(j, m, x)$ qui suppose l'existence de multiples prédicats lexicaux pour chaque verbe en fonction du nombre d'arguments qu'il rencontre dans un énoncé. On remarquera par ailleurs que la présence d'arguments facultatifs introduit des ambiguïtés potentielles d'interprétation (s'agit-il d'une localisation ou d'un élément temporel ?).

(3.12) *John kisses Mary in the plaza.*

(3.13) *She smiles.*

Le principe de Davidson nous permet de disposer d'une variable sur laquelle on applique de nouveaux prédicats, comme dans la représentation suivante de 3.12 :

$$\exists e.(Kiss(e) \wedge Ag(e, john) \wedge Pat(e, mary) \wedge Loc(e, plaza))\,^{2}$$

Dans ce cas, nous construisons la description d'un événement e qui est le fait d'embrasser, dans lequel il existe un patient et un agent, ainsi qu'une localisation. Il est relativement

2. *Ag* est utilisé pour Agent, *Pat* pour Patient et *Loc* pour Localisation

aisé de proposer une transformation des lexiques montagoviens pour produire ces représentations. L'un des avantages est que justement la structure argumentale du verbe n'est pas directement liée à son usage fonctionnel. On notera que le e utilisé ici pour les événements n'est pas le même que celui utilisé précédemment pour le contexte. Nous avons donc explicité comment utiliser ces concepts dans ce cadre, par exemple pour rendre compte des exemples 3.12 et 3.13 :

1. $[\![in\_the\_plaza]\!](([\![kiss]\!][\![Mary]\!])[\![John]\!])$
   $\Rightarrow_\beta \lambda eab.(Kiss(e) \wedge Ag(e, john) \wedge Pat(e, mary) \wedge Loc(e, plaza) \wedge b(e :: a))$
2. $[\![she]\!][\![smile]\!]$
   $\Rightarrow_\beta \lambda eab.(Smile(e) \wedge Ag(e, Sel(a)) \wedge b(e :: a))$

Il nous est apparu pertinent que l'intégration des événements dans ces représentations est assez proche de ce que la SDRT (Asher et Lascarides 2003) utilise pour produire des représentations complexes du discours. En SDRT il s'agit de construire des représentations des relations rhétoriques du discours. De notre côté, les variables d'événements introduites nécessitent d'être assemblées les unes avec les autres pour construire un tout cohérent. Nous avons donc proposé des solutions pour combiner les événements les uns aux autres, soit en utilisant une relation coordonnante, soit en utilisant une relation subordonnante comme le fait la SDRT.

Un élément en faveur de notre approche est que ces événements interviennent à des niveaux très différents. Au niveau de la phrase, ils sont utilisés pour construire leur représentation par agglomération de différentes informations, puis au niveau de la mise à jour du contexte où ils interviennent à un niveau de construction méta de la représentation.

Par ailleurs, Sai Qian a travaillé à la modélisation d'un autre phénomène linguistique : l'accessibilité des référents de discours sous l'effet de négations multiples. La difficulté est qu'en première approximation, comme nous l'avons introduit dans la section précédente, la continuation d'une négation est le contexte vide. Ainsi toute l'information récoltée à l'intérieur de la portée d'une négation est simplement perdue (voir exemple 3.14(b)-3.15), ce qui n'est pas toujours pertinent. Il arrive que nous utilisions plusieurs négations permettant de réouvrir l'accessibilité des référents de discours dans la portée d'une négation, comme dans l'exemple 3.14(c)-3.15.

(3.14) (a) Jean a une voiture
(b) Jean n'a pas de voiture
(c) Il n'est pas vrai que Jean n'a pas de voiture

(3.15) Elle est rouge

Sai Qian a intégré cette problématique dans la représentation. Pour cela, le contexte est composé d'un couple de positions qui contient simultanément une version et la version négative. La présence d'une négation a pour effet d'inverser les positions dans ce couple. Ainsi, sans modifier le type associé au discours qui reste Ω, les phrases sont interprétées avec le nouveau type :

$$[\![s]\!] = \Omega \times \Omega \tag{3.16}$$

Dans la paire ainsi construite la première composante correspond à la version positive et la seconde à la version négative de l'énoncé. L'incertitude sur ce qui sera retenu pour la suite de l'interprétation est uniquement conservé pendant le calcul de la représentation de la phrase. Le choix doit être réalisé au moment de la mise à jour du discours. Sai Qian introduit avec cette structure de couple des fonctions de projection sur chacune des composantes, π_1 qui extrait la première et π_2 la seconde. La négation échange les positions grâce à un terme nommé **swap** présenté dans l'équation 3.17.

$$\mathbf{swap} \triangleq \lambda A.\langle \pi_2 A, \pi_1 A\rangle \tag{3.17}$$

Ce terme a une propriété nécessaire à notre approche :

$$\mathbf{swap}(\mathbf{swap}\ M) = M \tag{3.18}$$

Cette modification structurelle du type a permis à Sai Qian de définir un nouveau cadre, DN-TTDL. On retrouvera l'ensemble des définitions et des preuves des propriétés du formalisme dans (Qian 2014b), en particulier les mécanismes qui permettent de dynamiser un lexique dans ce nouvel environnement.

Afin d'illustrer les principes de la proposition nous reprenons l'exemple 3.14(a)-3.15. Le résultat est une paire de termes, qui prennent en charge de manière différente la négation et la relation de contexte. Dans la première composante, les variables introduites (en rouge) sont passées à la continuation ce qui les rend accessibles, dans la seconde, seul le contexte gauche est passé au contexte doit.

$$\begin{aligned}[\![3.14(a)]\!]_{DN\text{-}TTDL} &= \overline{\overline{[\![have]\!]([\![a]\!][\![car]\!])[\![Jean]\!]}} \\ &\twoheadrightarrow_\beta \langle \lambda e\phi.(\exists x.(\mathbf{car}\ x \wedge \mathbf{have\ jean}\ x \wedge \phi(x :: e))), \\ &\qquad \lambda e\phi.(\neg(\exists x.(\mathbf{car}\ x \wedge \mathbf{have\ jean}\ x)) \wedge \phi e)\rangle\end{aligned}$$

Il en est de même pour 3.15 :

$$\begin{aligned}[\![3.15]\!]_{DN\text{-}TTDL} &= \overline{\overline{[\![is\_red]\!][\![it]\!]}} \\ &\twoheadrightarrow_\beta \langle \lambda e\phi.(\mathbf{red}\ (\mathsf{sel}\ e) \wedge \phi e), \lambda e\phi.(\neg(\mathbf{red}\ (\mathsf{sel}\ e)) \wedge \phi e)\rangle\end{aligned}$$

Ces deux représentations sont intégrées dans un contexte pour obtenir leur interprétation. Le terme met à jour le contexte avec le contenu sémantique de l'énoncé qui fixe l'interprétation en choisissant la première composante du couple. Dans notre exemple, la version qui introduit des référents de discours (en rouge) dans le contexte est celle qui est utilisée. Ainsi, nous pouvons calculer la forme suivante :

$$\begin{aligned}&= \mathbf{update}_{DN\text{-}TTDL}\ [\![D_{3.14(a)}]\!]_{DN\text{-}TTDL}\ [\![3.15]\!]_{DN\text{-}TTDL} \\ &\twoheadrightarrow_\beta \lambda e\phi.\exists x.(\mathbf{car}\ x \wedge \mathbf{have\ jean}\ x \wedge \mathbf{red}\ (\mathsf{sel}(x :: \mathbf{jean} :: \mathsf{nil}) \\ &\qquad\qquad \wedge \phi(x :: \mathbf{jean} :: \mathsf{nil}))\end{aligned}$$

où on constate que x est accessible pour le prédicat **red**. Comme nous le souhaitions, on note que :

$$\begin{aligned}[\![3.14(c)]\!]_{DN\text{-}TTDL} &= \overline{\overline{[\![not]\!]([\![not]\!]([\![have]\!]([\![a]\!][\![car]\!])[\![Jean]\!]))}} \\ &= \overline{([\![have]\!]([\![a]\!][\![car]\!])[\![Jean]\!])}\end{aligned}$$

Dans la version négative de 3.14(b) le calcul produit la représentation suivante :

$$\begin{aligned}[\![3.14(b)]\!]_{DN\text{-}TTDL} &= \overline{\overline{[\![not]\!]([\![have]\!]([\![a]\!][\![car]\!])[\![Jean]\!])} \\ &= \neg^{d}_{DN\text{-}TTDL}\overline{\overline{([\![have]\!]([\![a]\!][\![car]\!])[\![Jean]\!])}} \\ &\twoheadrightarrow_{\beta} \langle \lambda e\phi.(\neg(\exists x.(\mathbf{car}\ x \wedge \mathbf{have\ jean}\ x)) \wedge \phi e), \\ &\qquad \lambda e\phi.(\exists x.(\mathbf{car}\ x \wedge \mathbf{have\ jean}\ x \wedge \phi(x :: e)))\rangle\end{aligned}$$

Cette dernière est également combinée avec la suite du discours, 3.15. Dans ce cas, la sélection n'a pas accès au référent de discours qui dénote de la voiture. Il n'est donc pas possible de résoudre correctement l'anaphore pronominale.

$$\begin{aligned}&= \mathbf{update}_{DN\text{-}TTDL}\ [\![D_{3.14(b)}]\!]_{DN\text{-}TTDL}\ [\![3.15\text{-}2]\!]_{DN\text{-}TTDL} \\ &\twoheadrightarrow_{\beta} \lambda e\phi.((\neg(\exists x.(\mathbf{car}\ x \wedge \mathbf{have\ jean}\ x))) \wedge \mathbf{red}\ (\mathsf{sel}(\mathbf{jean} :: \mathsf{nil})) \\ &\qquad\qquad \wedge \phi(\mathbf{jean} :: \mathsf{nil}))\end{aligned}$$

Si les exemples qui utilisent explicitement la double négation peuvent sembler relativement rares, ce n'est pas le cas des interactions entre les connecteurs logiques qui peuvent faire apparaître des négations implicites. Un exemple proposé par (Roberts 1989) (et attribué à Barbara Partee) est repris dans l'exemple 3.19.

(3.19) *Either there is no bathroom in this appartement or it is in funny place.*
Soit il n'y a pas de baignoire dans cet appartement, soit elle est dans un endroit bizarre.

La disjonction est construite autour d'un élément qui contient une négation, ce qui produit : $\neg\phi \vee \psi = \neg(\neg(\neg\phi) \wedge \neg\psi)$. On retrouve également ce type d'interaction dans les négations implicites dont on trouve une illustration dans l'exemple 3.20.

(3.20) *A student$_i$ passed the examination. He$_i$ studied hard.*
Un étudiant a réussi l'examen. Il avait beaucoup travaillé.
*Not every student$_i$ failed the examination. *He$_i$ studied hard.*
Tous les étudiants n'ont pas échoué à l'examen. * Il a beaucoup travaillé.

3.3 Subordination modale

La dernière partie du travail de Sai Qian a été de travailler sur la modélisation de la subordination modale (Hintikka 1957 ; von Fintel 2006 ; Hacquard 2006). Les motivations sont proches de celles du problème des négations multiples car les modalités ont également la capacité d'ouvrir à nouveau l'accessibilité des référents de dicours, comme montré dans l'exemple 3.21.

(3.21) Jean n'a pas de voiture.
Il ne saurait pas où la garer.

La subordination modale est le phénomène par lequel un groupe nominal dans la portée d'un opérateur modal peut servir d'antécédent pour une expression anaphorique qui apparaît dans un autre énoncé, comme c'est le cas dans les exemples 3.22, 3.23 et 3.24 issus de (Roberts 1989).

(3.22) *If John bought a book$_i$, he'll be home reading it$_i$ by now. It$_i$ 'll be a murder mystery.*

(3.23) *A thief$_i$ might break into the house. He$_i$ would take the silver.*

(3.24) *A wolf$_i$ walks in. It$_i$ might growl.*

Sai Qian s'est pour cela inspiré de la théorie des modalités développée par Angelika Kratzer (Kratzer 1977 ; Kratzer 1981 ; Kratzer 1986 ; Kratzer 1991). Ces théories ont largement été reprises pour la question de la formalisation de la subordination modale. Dans la théorie de Kratzer, l'interprétation des modaux dépend de leur contexte. Par ailleurs, les modaux sont la nécessité, notée $\Box$, et la possibilité, notée $\Diamond$. Le contexte contient de fait les propriétés vraies à ce point de l'interprétation, autrement dit, dans ce monde d'interprétation. On définit alors la notion de *conversational background* qui est une fonction des mondes possibles vers des ensembles de propositions (modales).

Cette approche a l'avantage de réduire le problème de l'interprétation de la modalité à celui de la relation d'accessibilité qui est déterminée par la notion de *conversational background*. Pour f un *conversational background*, w, u des variables qui représentent des mondes possibles, u est accessible depuis w en fonction de f, avec la notation $\mathbf{R}_f(w, u)$, si et seulement si $u \in [\![\wedge f(w)]\!]^M_{MPL}$, ou, si et seulement si toutes les propositions de $f(w)$ sont vraies dans u. Par conséquent, l'ensemble des mondes qui sont accessibles depuis w en fonction de f est $[\![\wedge f(w)]\!]^M_{MPL}$.

Une part importante des travaux de Kratzer sur la modalité met en avant les propriétés formelles entre l'ensemble des mondes possibles W, la fonction d'interprétation I, le *conversational background* f et la relation d'accessibilité R_f (sérialité, réflexivité, transitivité, identité pour la relation d'accessibilité et consistance, réalité, rétrospective positive pour le *conversational background*).

De nombreux travaux ont étudié la modélisation des modalités (Sells 1985 ; Roberts 1987 ; Roberts 1989 ; Van Rooij 2005). Une première solution dans TTDL a été proposée dans (Asher et Pogodalla 2010). Pour traiter le problème, ils introduisent des opérateurs modaux et une nouvelle structure pour le contexte. Ce dernier contient alors un ensemble d'entités, comme dans les usages traditionnels du contexte, et une base modale (c'est-à-dire l'ensemble des propositions nécessairement vraies). Leur proposition rend fidèlement compte de (Roberts 1989) dans la théorie des continuations, mais les usages croisés de nécessité et de possibilité nécessitent une représentation plus fine des interactions possibles. La solution de Sai Qian suit une autre perspective où les mondes possibles sont représentés par des variables spécifiques.

De plus, la notion de contexte évolue pour se rapprocher de l'idée de base modale. Pour y parvenir, Sai Qian ajoute à la logique un nouveau quantificateur existentiel qui

porte sur des mondes possibles et non sur des individus : ${}^{s}\exists$. Une nouvelle constante qui dénote du monde courant dans lequel l'interprétation est réalisée est également nécessaire, notée **H** (pour *here*).

Clarifions une modification importante : dans un système modal, la recherche de valeurs de vérité n'a pas de sens, puisque l'interprétation dépend justement du monde utilisé pour l'interprétation. Nous nous référerons donc à des ensembles de mondes possibles dans lesquels les propositions sont vraies. Les types ne seront donc plus seulement o mais $o_i = s \to o$ (monde possible vers valeur de vérité).

Pour expliquer l'interprétation faite des contextes, la notion d'environnement est introduite. Il s'agit ici d'une paire contenant le *background* et les informations de base. Le background contient les propositions vraies dans un monde donné et l'autre composante permet de transmettre des propositions mises à jour des mondes possibles vers les mondes accessibles. Les environnements sont donc du type $o_i \times o_i$.

À partir des environnements, Sai Qian introduit les environnements généralisés qui sont le pendant des *conversational backgrounds* de Kratzer. Dans ce cas, il s'agit d'une projection des mondes possibles vers les environnements. En utilisant un environnement généralisé à un monde particulier, l'environnement de ce monde est donc $T_{genv} = s \to T_{env}$

Même si nous ne l'avons pas encore explicité, le principe de cette modélisation est de positionner des informations vraies dans certains mondes en fonction de la relation d'accessibilité entre ces mondes. La notion de monde d'accès (*world of interest*) sert à enregistrer la position actuelle du calcul par rapport à l'ensemble de la représentation des mondes. Une subtilité importante est qu'il ne s'agit pas là du monde dans lequel un énoncé est interprété, mais bien de celui à partir duquel il se positionne pour l'interprétation. L'introduction de ce monde d'accès permet de définir le type des contextes gauches.

$$\begin{aligned} \gamma_i &\triangleq s \times T_{genv} \\ &= s \times (s \to ((s \to o) \times (s \to o))) \end{aligned} \tag{3.25}$$

En reportant ces modifications de type, l'interprétation du contexte droit, qui est une fonction des contextes gauches vers les propositions modales, est obtenue : $\gamma_i \to o_i = (s \times T_{genv}) \to o_i$ à partir de laquelle le type des phrases et des discours est défini :

$$[\![s]\!] = \gamma_i \to (\gamma_i \to o_i) \to o_i \tag{3.26}$$

$$[\![d]\!] = \gamma_i \to (\gamma_i \to o_i) \to o_i \tag{3.27}$$

Grâce à ces définitions, Sai Qian redéfinit les connecteurs logiques utilisés pour construire les représentations, en particulier, pour introduire la notion de modalité, ainsi que les termes permettant de manipuler les mondes et de passer de l'un à l'autre :

— conjonction modale $\wedge_i \triangleq \lambda ABi.(Ai \wedge Bi) : o_i \to o_i \to o_i$
— négation modale $\neg_i \triangleq \lambda Ai.\neg(Ai) : o_i \to o_i$
— quantificateur existentiel modal pour les individus ${}^{\iota}\exists_i \triangleq \lambda Pi.{}^{\iota}\exists(\lambda x.Pxi) : (\iota \to o_i) \to o_i$
— tautologie modale $\top_i \triangleq \lambda i.\top : o_i$

Sai Qian ajoute plusieurs termes permettant de manipuler des mondes d'accès et des environnements :

— retrouver le monde d'accès (*world of interest*) : $\gamma_i \to s$

$$\textbf{woi} \triangleq \lambda e.\pi_1 e \tag{3.28}$$

La fonction **woi** est relativement directe. Elle prend un contexte gauche e en entrée et retourne son monde d'accès, qui est simplement la première projection de e.

— retrouver l'environnement généralisé : $\gamma_i \to T_{genv}$

$$\textbf{genv} \triangleq \lambda e.\pi_2 e \tag{3.29}$$

La fonction **genv** prend un contexte gauche e et retourne son environnement généralisé, qui est la seconde composante de e.

— retrouver l'environnement : $\gamma_i \to s \to T_{env}$

$$\textbf{env} \triangleq \lambda ei.(\textbf{genv}\ e\ i) \tag{3.30}$$

La fonction **env** est construite à partir de **genv** (formule 3.29). Elle prend un contexte gauche et un monde possible, et retourne un environnement spécifique pour le monde d'entrée.

— modifier le monde d'accès : $\gamma_i \to s \to \gamma_i$

$$\textbf{change\_woi} \triangleq \lambda ei.\langle i, (\textbf{genv}\ e)\rangle \tag{3.31}$$

La fonction **change_woi** prend un contexte gauche e et un monde possible i en entrée. Elle produit un nouveau contexte gauche, où le monde d'accès est modifié avec le monde passé en paramètre i, l'environnement généralisé est celui du contexte gauche d'entrée.

— retrouver le *background* : $\gamma_i \to s \to o_i$

$$\textbf{bkgd} \triangleq \lambda ei.\pi_1(\textbf{env}\ e\ i) \tag{3.32}$$

La fonction **bkgd** prend un contexte gauche e (un produit cartésien composé d'un monde d'accès et d'un environnement généralisé) et un monde possible i en entrée. Elle produit une proposition modale, qui est le *background* (le premier élément de l'environnement) du contexte gauche e pour le monde i.

— retrouver la base : $\gamma_i \to s \to o_i$

$$\textbf{base} \triangleq \lambda ei.\pi_2(\textbf{env}\ e\ i) \tag{3.33}$$

La fonction **base** prend un contexte gauche e (un produit cartésien composé d'un monde d'accès et d'un environnement généralisé) et un monde possible i en entrée. Elle produit une proposition modale, qui est la base (le second élément de l'environnement) du contexte gauche e dans le monde i.

Enfin, les termes nous permettant de manipuler les contextes à partir des termes précédents sont introduits. Seules les premières définitions sont reprises ici pour dresser les perspectives. L'ensemble de ces définitions est disponible dans (Qian 2014b ; Qian, de Groote et Amblard 2016).

— La fonction **up_genv** met à jour un environnement généralisé avec une nouvelle valeur de monde d'accès.

$$\textbf{up\_genv} \triangleq \lambda GiE.G[i := E] \tag{3.34}$$

— La fonction **up_context** met à jour un contexte en y incluant une nouvelle proposition modale vraie à la fois dans le *balckground* et dans la base.

$$\begin{aligned}\textbf{up\_context} \triangleq \lambda eiA.\langle &(\textbf{woi}\ e),\\ &\textbf{up\_genv}\\ &\qquad (\textbf{genv}\ e)\\ &\qquad i\\ &\qquad \langle A \wedge_i (\textbf{bkgd}\ e\ i), A \wedge_i (\textbf{base}\ e\ i)\rangle\\ &\rangle\end{aligned} \tag{3.35}$$

— La fonction **copy_context** copie l'environnement d'un monde dans celui d'un autre monde.

— La fonction **reset_base** produit à partir d'un contexte gauche et d'un monde possible, un nouveau contexte gauche qui a toujours le même monde d'accès, mais la base de l'environnement généralisé est remise à zéro.

— le contexte gauche vide.

$$\mathsf{nil}_i \triangleq \langle \mathbf{H}, \lambda i.\langle \top_i, \top_i\rangle\rangle \tag{3.36}$$

— et le contexte droit vide.

$$\textbf{stop}_i \triangleq \lambda e.\top_i \tag{3.37}$$

Ces définitions étant techniques, il est difficile d'avoir une vision globale du processus de calcul *a priori*. Les éléments précédemment introduits sont les briques de base pour écrire les termes associés aux éléments lexicaux. En particulier, Sai Qian définit les opérateurs modaux traditionnels.

L'opérateur modal de possibilité prend une proposition dynamique A en entrée et construit une nouvelle proposition dynamique $\Diamond A$ pour laquelle il existe un nouveau monde accessible qui vérifie A. Comme les propositions de la base sont vraies dans tous les mondes accessibles, elles le sont dans ce nouveau monde. Le contexte est également transféré vers ce nouveau monde. Enfin, le monde d'accès est mis à jour avec ce nouveau monde :

$$\begin{aligned}\Diamond \triangleq \lambda Ae\phi i.{}^s\exists j.(&\mathbf{R}\ i\ j \wedge\\ &\textbf{base}\ e\ i\ j \wedge\\ &A\ (\textbf{copy\_context}\ e\ i\ j)\\ &\quad (\lambda e'j'.\phi\ (\textbf{reset\_base}\ (\textbf{change\_woi}\ e'\ j')i)\ i\ )\\ &\quad j)\end{aligned} \tag{3.38}$$

La figure 3.4 présente graphiquement des relations entre les mondes pour A et pour $\Diamond A$. Les lignes pointillées représentent des liens d'accessibilité non spécifiés, contrairement aux lignes pleines. Le monde d'accès est celui en rouge et mb est la fonction qui associe la base modale.

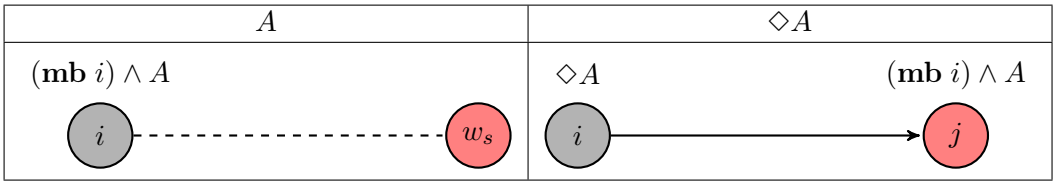


FIGURE 3.4 – Relations entre les mondes possibles pour $\Diamond$

L'opérateur de nécessité prend une proposition A et construit une proposition modale $\Box A$. La base modale du monde d'accès est copiée dans tous les mondes accessibles. Mais les référents de discours introduits dans la portée de la modalité ne sont pas transmis en dehors. **Stop** est alors passé comme contexte droit, ce qui arrête le processus de continuation. Comme pour la modalité précédente, Sai Qian représente graphiquement les relations dans la figure 3.5.

$$\begin{aligned}\Box \triangleq \lambda A e \phi i.(^{s}\forall j.(&\mathbf{R}\ i\ j \rightarrow \\ &(\mathbf{base}\ e\ i\ j \rightarrow \\ &(A\ (\mathbf{copy\_context}\ e\ i\ j)\ \mathbf{stop}_i\ j)))) \\ &\wedge \phi e i\end{aligned} \tag{3.39}$$

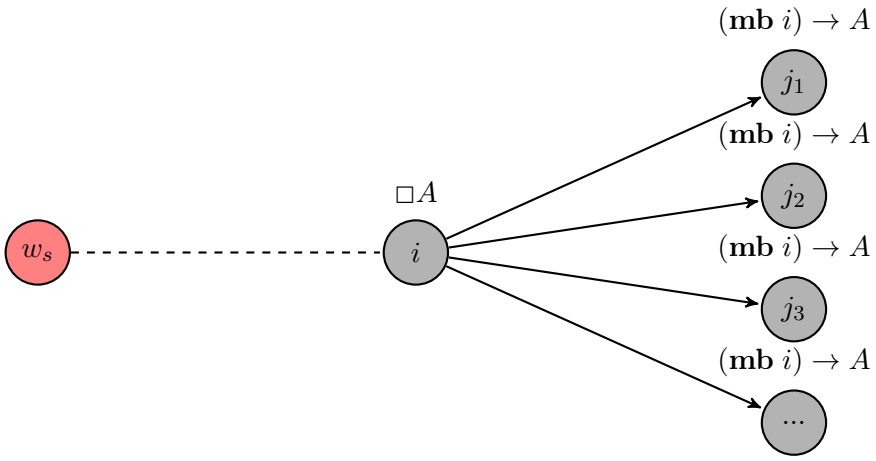


FIGURE 3.5 – Relation entre les mondes possibles de $\Box A$

Il reste un dernier élément à introduire qui est un terme permettant d'interpréter une proposition dans un monde particulier. On note **at** cette entrée, qui permet de définir

par exemple les entrées lexicales des modaux :

$$[\![might]\!]_{M\text{-}TTDL} = \lambda A e \phi i.(\mathbf{at}\ (\mathbf{woi}\ e)\ (\Diamond A))e\phi i \tag{3.40}$$

$$[\![would]\!]_{M\text{-}TTDL} = \lambda A e \phi i.(\mathbf{at}\ (\mathbf{woi}\ e)\ (\Box A))e\phi i \tag{3.41}$$

Des illustrations peuvent être trouvées dans (Qian 2014b ; Qian, de Groote et Amblard 2016), en particulier le traitement complet de l'exemple 3.24, ainsi que la définition du processus automatique de dynamisation des lexiques dans ce cadre.

Chapitre 4

Sémantique de la langue par les effets algébriques

Sommaire

Dans la partie précédente nous avons travaillé à rendre compte de phénomènes fins en mobilisant la flexibilité de notre formalisme et en maintenant la cohérence de la proposition grâce à la compositionalité de l'approche. S'il est possible de proposer des traitements particuliers, une question importante est celle de rassembler ces propositions dans une unique grammaire.

Nous avons alors travaillé avec Jirka Maršík à la définition du problème de composition de plusieurs grammaires ACG dans (Maršík et Amblard 2013). Le résultat obtenu n'étant pas suffisant pour envisager un traitement à large couverture, Jirka Maršík a proposé dans sa thèse, co-encadrée avec Philippe de Groote, d'utiliser les effets algébriques (*effects*) et *handlers* en les intégrant dans le λ-calcul simplement typé avec une monade. Ces théories ont connu de nombreux développements ces dernières années, et permettent un transfert relativement immédiat des concepts utilisés dans des langages de programmation proches de ceux que nous utilisons, comme Haskell.

Hyland, Power et Plotkin ont étudié le problème de dériver la sémantique dénotationnelle des langages de programmation qui combinent différents effets algébriques (Hyland, G. Plotkin et Power 2006). Plutôt que de modéliser les effets par des monades et de combiner les monades, les effets sont définis par des opérateurs sur des calculs qui sont alors des expressions algébriques avec des effets (des opérations) et des valeurs. Pour combiner deux ensembles, les signatures sont additionnées et les théories sont combinées. Afin de modéliser les exceptions, Plotkin et Pretnar ont enrichi la théorie avec la notion de *handler* (G. D. Plotkin et Pretnar 2013). L'objectif d'un *handler* est de remplacer les occurrences d'un opérateur dans un calcul par une autre expression.

Le calcul ainsi obtenu permet de simuler le passage de propriétés à l'extérieur de calculs spécifiques, ou au contraire contraint d'interpréter un phénomène à un moment particulier de l'évaluation. Ainsi, Jirka Maršík définit un calcul basé sur le λ-calcul qui simule des stratégies d'évaluation différentes, en particulier les appels par valeur [1] (*call by value*) ou les appels par nom [2] (*call by name*). Jirka Maršík a nommé ce calcul le

1. Certainement la stratégie la plus courante, qui consiste à évaluer les arguments d'une fonction avant de les passer à la fonction.

2. Dans cette stratégie d'évaluation les arguments d'une fonction ne sont évalués qu'au moment où

λ-banane, noté $(\!|\lambda|\!)$.

4.1 Composition des traitements

Dans l'objectif de constituer des grammaires de type ACG à large couverture Jirka Maršík a proposé - dans le cadre de son master - une extension des ACG qui permet d'exprimer des contraintes, dans des signatures différentes, qui s'appliquent sur une même autre signature (Maršík 2013). Ainsi, la structure traditionnelle d'arbre des ACG est abandonnée pour celle de DAG. L'origine de la proposition vient du caractère typé des ACG. Si l'on décide d'inclure le traitement d'un trait particulier dans la grammaire, cela a un impact direct sur le type. Par exemple, l'un des 12 types associés à la préposition *de* qui exprime en même temps des propriétés de négation, de nombre et de genre peut être :

$$C_{de_{11}} : (NP\_NEG{=}T\_VAR{=}F\_NUM{=}PL) \multimap (N\_NEG{=}F\_NUM{=}SG) \multimap (N\_NEG{=}T\_NUM{=}SG)$$

La figure 4.1 reprend les structures hiérarchiques entre les différentes signatures (par les lexiques). En (a), on peut observer que l'ajout d'un système de contraintes doit se faire en conservant la structure d'arbre. De fait, le système de contraintes n'est pas introduit pour lui-même, mais par une meta-signature qui le contient. De l'autre côté, en 4.1(b), la structure des G-ACG permet de faire porter plusieurs systèmes de contraintes sur une même signature, avec partage des langages engendrés.

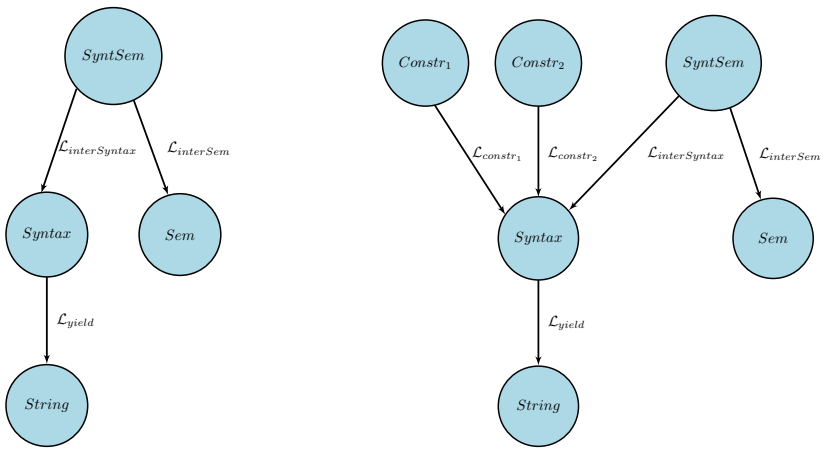


FIGURE 4.1 – Représentation d'une ACG et d'une G-ACG avec partage de contraintes

ils sont appelés (avec leur nom). Les arguments non utilisés ne sont alors pas évalués, mais un argument utilisé plusieurs fois le sera plusieurs fois.

Les *graphical abstract categorial grammars* (G-ACGs) généralisent les ACG. Il est alors possible de définir différents langages reconnus par les G-ACG en fonction de ceux des ACG (*abstract and object languages*, *extrinsic languages*, *pangraphical languages*, *etc.*). Jirka Maršík a montré comment les G-ACGs permettent d'écrire des grammaires qui capturent les différents traitements de phénomènes sémantiques (Maršík et Amblard 2013).

Avec l'exemple, on entrevoit le risque d'explosion du nombre de types associés à chaque mot dans le lexique. Or les complexités d'analyse étant dépendantes de ce nombre d'éléments, l'efficacité est limitée. Par ailleurs, la gestions des interactions entre les différentes contraintes de types possibles est trop complexe.

4.2 Définitions et propriétés du $(\!|\lambda|\!)$

Pour définir le $(\!|\lambda|\!)$, un constructeur de type $\mathcal{F}$ est introduit dans le langage. Le type $\mathcal{F}(\alpha)$ correspond à un calcul qui produit une valeur de type α. Ce principe est inspiré des langages de programmation qui utilisent les monades (Moggi 1991 ; Wadler 1992 ; Jones 2003). Le passage de α à $\mathcal{F}(\alpha)$ peut être vu comme une généralisation souvent utilisée en sémantique, par exemple le *type raising* (Montague 1973b), la dynamisation (Lebedeva 2012), l'intensionnalisation (de Groote et Kanazawa 2013). La différence pour expliciter la relation entre interprétation sémantique au niveau de α et possibilité d'interprétation pragmatique pour $\mathcal{F}(\alpha)$ a été abordée dans (Maršík et Amblard 2015).

Le calcul dispose d'un constructeur qui peut combiner tous les autres, en particulier au niveau des monades par des transformations de monades. Cette technique a été introduite par Moggi (Moggi 1991) et reprise pour la modélisation de la langue naturelle par Simon Charlow (Charlow 2014).

Dans le langage, les valeurs du type $\mathcal{F}(\alpha)$ sont vues comme des expressions algébriques construites à partir de signatures d'effets et d'un générateur d'ensemble α. Comme une expression algébrique est soit une constante, soit un opérateur appliqué à une autre expression, nous pouvons utiliser les effets de bord. Un calcul est alors soit une valeur, soit un effet couplé à une continuation.

Les valeurs de type α sont projetées vers le type des expressions algébriques $\mathcal{F}(\alpha)$. Il suffit de construire des expressions utilisant des opérateurs sur la signature d'effets, autrement dit de créer des calculs qui appellent la réalisation de cet effet.

L'ensemble des définitions est présenté dans (Maršík 2016), mais nous revenons ici sur les idées principales. La construction des termes du langage se fait sur plusieurs ensembles : $\mathcal{X}$ un ensemble de variables, Σ une signature typée et $\mathcal{E}$ un ensemble de symboles d'opérations. Les expressions du langage sont constituées des éléments suivants (les quatre premières constructions proviennent du λ-calcul simplement typé et les quatre autres des expressions algébriques pour réaliser le calcul).

abstraction $\lambda x.\, M$, où x est une variable de $\mathcal{X}$ et M est une expression

application $M\, N$, où M et N sont des expressions

variable x, où x est une variable de $\mathcal{X}$

$$\frac{\Gamma, x : \alpha \vdash M : \beta}{\Gamma \vdash \lambda x.\, M : \alpha \to \beta}\ [\text{abs}] \qquad \frac{\Gamma \vdash M : \alpha \to \beta \qquad \Gamma \vdash N : \alpha}{\Gamma \vdash M\, N : \beta}\ [\text{app}]$$

$$\frac{x : \alpha \in \Gamma}{\Gamma \vdash x : \alpha}\ [\text{var}] \qquad \frac{c : \alpha \in \Sigma}{\Gamma \vdash c : \alpha}\ [\text{const}]$$

$$\frac{\Gamma \vdash M : \alpha}{\Gamma \vdash \eta\, M : \mathcal{F}_E(\alpha)}\ [\eta] \qquad \frac{\Gamma \vdash M_p : \alpha \qquad \Gamma, x : \beta \vdash M_c : \mathcal{F}_E(\gamma) \qquad \mathtt{op} : \alpha \rightarrowtail \beta \in E}{\Gamma \vdash \mathtt{op}\, M_p\, (\lambda x.\, M_c) : \mathcal{F}_E(\gamma)}\ [\mathtt{op}]$$

$$\frac{\Gamma \vdash M : \mathcal{F}_\emptyset(\alpha)}{\Gamma \vdash \mathbin{⫰} M : \alpha}\ [⫰] \qquad \frac{\begin{array}{c} E = \{\mathtt{op}_i : \alpha_i \rightarrowtail \beta_i\}_{i \in I} \uplus E_f \\ E' = E'' \uplus E_f \\ [\Gamma \vdash M_i : \alpha_i \to (\beta_i \to \mathcal{F}_{E'}(\delta)) \to \mathcal{F}_{E'}(\delta)]_{i \in I} \\ \Gamma \vdash M_\eta : \gamma \to \mathcal{F}_{E'}(\delta) \\ \Gamma \vdash N : \mathcal{F}_E(\gamma) \end{array}}{\Gamma \vdash (\!|\, (\mathtt{op}_i \colon M_i)_{i \in I},\ \eta \colon M_\eta \,|\!)\, N : \mathcal{F}_{E'}(\delta)}\ [(\!|\,|\!)]$$

$$\frac{\Gamma \vdash M : \alpha \to \mathcal{F}_E(\beta)}{\Gamma \vdash \mathcal{C}\, M : \mathcal{F}_E(\alpha \to \beta)}\ [\mathcal{C}]$$

FIGURE 4.2 – Règles de typage pour le $(\!|\, \lambda \,|\!)$

constante c, où c est une variable de Σ

operation $\mathtt{op}\, M_p\, (\lambda x.\, M_c)$, où $\mathtt{op}$ est un opérateur de $\mathcal{E}$, x est une variable de $\mathcal{X}$ et, M_p et M_c sont des expressions

injection $\eta\, M$, où M est une expression

handler $(\!|\, \mathtt{op}_1 \colon M_1,\ \ldots,\ \mathtt{op}_n \colon M_n,\ \eta \colon M_\eta \,|\!)$ où les $\mathtt{op}_i$ sont des opérateurs de $\mathcal{E}$ et, M_i et M_η des expressions

extraction ⫰

échange $\mathcal{C}$

On suppose que la fonction η retourne la valeur donnée en paramètre. Les parenthèses bananes $(\!|\, \mathtt{op}_1 \colon M_1,\ \ldots,\ \mathtt{op}_n \colon M_n,\ \eta \colon M_\eta \,|\!)$ décrivent les *handlers* : elles contiennent les clauses qui permettent d'arrêter un calcul à partir des opérations. La fonction cerise ⫰ s'applique à un calcul qui ne contient pas d'effets et le réalise pour produire son résultat.

Les règles de typage sont présentées dans la figure 4.2. Les métavariables M, $N\ldots$ désignent les expressions, α, β, $\gamma\ldots$ les types, Γ, $\Delta\ldots$ les contextes, $\mathtt{op}$, $\mathtt{op}_i$ les symboles d'opérations et E, $E'\ldots$ les signatures d'effets. Σ fait référence à la signature d'ordre supérieur qui fournit les types des constantes.

$(\lambda x.\, M)\, N \to$ $M[x := N]$	règle β
$\lambda x.\, M\, x \to$ M	règle η où $x \notin \mathrm{FV}(M)$
$(\!|\, (\mathtt{op}_i \colon M_i)_{i \in I},\ \eta \colon M_\eta \,|\!)\, (\eta\, N) \to$ $M_\eta\, N$	règle $(\!|\,\eta\,|\!)$
$(\!|\, (\mathtt{op}_i \colon M_i)_{i \in I},\ \eta \colon M_\eta \,|\!)\, (\mathtt{op}_j\, N_\mathrm{p}\, (\lambda x.\, N_\mathrm{c})) \to$ $M_j\, N_\mathrm{p}\, (\lambda x.\, (\!|\, (\mathtt{op}_i \colon M_i)_{i \in I},\ \eta \colon M_\eta \,|\!)\, N_\mathrm{c})$	règle $(\!|\,\mathtt{op}\,|\!)$ où $j \in I$ et $x \notin \mathrm{FV}((M_i)_{i \in I}, M_\eta)$
$(\!|\, (\mathtt{op}_i \colon M_i)_{i \in I},\ \eta \colon M_\eta \,|\!)\, (\mathtt{op}_j\, N_\mathrm{p}\, (\lambda x.\, N_\mathrm{c})) \to$ $\mathtt{op}_j\, N_\mathrm{p}\, (\lambda x.\, (\!|\, (\mathtt{op}_i \colon M_i)_{i \in I},\ \eta \colon M_\eta \,|\!)\, N_\mathrm{c})$	règle $(\!|\,\mathtt{op}'\,|\!)$ où $j \notin I$ et $x \notin \mathrm{FV}((M_i)_{i \in I}, M_\eta)$
$\downarrow\!\!\circ\, (\eta\, M) \to$ M	règle $\downarrow\!\!\circ$
$\mathcal{C}\, (\lambda x.\, \eta\, M) \to$ $\eta\, (\lambda x.\, M)$	règle $\mathcal{C}_\eta$
$\mathcal{C}\, (\lambda x.\, \mathtt{op}\, M_\mathrm{p}\, (\lambda y.\, M_\mathrm{c})) \to$ $\mathtt{op}\, M_\mathrm{p}\, (\lambda y.\, \mathcal{C}\, (\lambda x.\, M_\mathrm{c}))$	règle $\mathcal{C}_\mathtt{op}$ où $x \notin \mathrm{FV}(M_\mathrm{p})$

FIGURE 4.3 – Règles de réduction pour le λ-banane

À partir de ces règles, une sémantique est associée au calcul qui est donnée par une relation de réduction sur les termes. Comme le calcul se focalise sur la notion d'effets, il n'a pas de notion d'ordre d'évaluation, toute sous-expression réductible peut l'être dans n'importe quel contexte. À partir de ces éléments plusieurs concepts sont définis comme la notion de variable libre sur les termes, de substitution, ainsi que les nouvelles variables qui sont des variables dont le nom n'a pas encore été utilisé. Les règles de réduction sont présentées dans la figure 4.3.

À partir de ces règles de typage et de réductions, nous pouvons contrôler dans le calcul l'ordre d'évaluation et le résultat obtenu. Il est donc possible de simuler des appels par valeur et des appels par nom. En général, les *handlers* sont utilisés pour des usages particuliers comme la gestion des erreurs, mais ici, ils sont au centre du calcul. Ils permettent de récupérer des évaluations en train de remonter à travers des appels successifs. Ainsi, ils introduisent des points singuliers dans l'analyse. En terme sémantique, il est

possible de faire le lien entre portée et *handlers*. La portée peut ne pas être définie *a priori*, elle est alors rattrapée et fixée par un *handler*.

Ce calcul une fois défini, il faut montrer comment le mobiliser pour simuler la sémantique. De nombreux exemples ont été développés et sont disponibles dans (Maršík 2016). Pour justifier l'approche, il était nécessaire que ce calcul ne diverge pas. Jirka Maršík a prouvé les propriétés importantes : la préservation de type qui assure que l'application des règles sur une expression ne modifie pas son type. Il y a une garantie de stabilité dans l'application du calcul.

Il a par ailleurs prouvé la confluence du système. Cette propriété assure qu'un terme ne peut pas être réécrit sous deux formes différentes qui conduiraient à des résultats différents. Ainsi lorsqu'un choix d'application de règles se présente, le choix de l'une par rapport à l'autre implique qu'il existe une autre règle à appliquer qui conduit au même résultat. Pour parvenir à cette preuve Jirka Maršík montre que le $(\!|\,\lambda\,|\!)$ peut être modélisé en terme de Combinatory Reduction Systems (CRS) de (Klop, Van Oostrom et Van Raamsdonk 1993) qui possède la propriété de confluence. La difficulté provient de la règle η qui supprime l'orthogonalité des CRS (et donc la confluence). Il s'agit de montrer que $(\!|\,\lambda\,|\!)$ sans la règle η est un CRS et que la règle η commute avec toutes les autres. Le calcul est divisé en deux parties, l'une sans règle η qui a la propriété, et celle qui contient toutes les applications des règles η (et qui est confluente). Le $(\!|\,\lambda\,|\!)$ est donc lui-même confluent.

La même stratégie est utilisée pour prouver la terminaison en utilisant les Inductive Data Type Systems (IDTS) (Blanqui 2000 ; Blanqui, Jouannaud et Okada 2002). Comme les CRS, IDTS sont des systèmes de réécriture pour lesquels certaines propriétés générales sont prouvées. La stratégie est donc de montrer la correspondance de $(\!|\,\lambda\,|\!)$ et IDTS. IDTS possède une condition suffisante pour la terminaison appelée *general scheme*. $(\!|\,\lambda\,|\!)$ ne satisfait malheureusement pas cette condition, mais Jirka Maršík a montré comment le transformer en suivant les techniques d'étiquetage sémantique d'ordre supérieur de Hamana (Hamana 2007). Comme pour la confluence, le $(\!|\,\lambda\,|\!)$ sans la règle η est d'abord étudié, avant d'y ajouter ces règles en préservant la terminaison.

Pour montrer la flexibilité du calcul, Jirka Maršík a introduit une simulation des appels par valeur et des appels par nom. Pour cela, la réduction d'un terme en une étape est définie par : $C[(\lambda x.\, M)\, V] \to_\beta C[M[x := V]]$.

Pour réaliser la simulation de l'appel par valeur nous avons projeté λ_v vers $(\!|\,\lambda\,|\!)$. Pour M un terme de λ_v, son interprétation dans $(\!|\,\lambda\,|\!)$, soit $[\![M]\!]$, est :

$$\begin{aligned} [\![x]\!] &= \eta\, x \\ [\![\lambda x.\, M]\!] &= \eta\, (\lambda x.\, [\![M]\!]) \\ [\![M\, N]\!] &= [\![M]\!] \gg\!= (\lambda m.\, [\![N]\!] \gg\!= (\lambda n.\, m\, n)) \end{aligned}$$

$\gg\!=$ est un constructeur (*binder*). $M \gg\!= N$ est alors le programme qui exécute M pour obtenir le résultat x et poursuit avec le programme $N\, x$.

Le calcul permet de simuler le λ-calcul simplement typé dans $(\!|\,\lambda\,|\!)$ (qui est typé), et également de simuler λ_v non-typé dans une version non-typée de $(\!|\,\lambda\,|\!)$. Cela peut être illustré par l'introduction des opérateurs de contrôle `shift0` et `reset0`. En fait, les effets

et les *handlers* sont très proches des continuations délimitées. Cette paire d'opérateurs est celle qui se rapproche le plus du comportement des *handlers*. Pour simuler $\lambda_{\texttt{shift0}}$ dans $(\!|\,\lambda\,|\!)$, la simulation de λ_v est étendue avec les interprétations des deux nouvelles formes syntaxiques :

$$[\![\texttt{shift0}\,M]\!] = [\![M]\!] \gg\!= (\lambda m.\,\texttt{shift0!}\,m)$$
$$[\![\texttt{reset0}\,M]\!] = (\!|\,\texttt{shift0:}\,(\lambda ck.\,c\,k)\,|\!)\,[\![M]\!]$$

Ces égalités nous donnent un modèle général pour simuler le calcul avec des effets de bord dans $(\!|\,\lambda\,|\!)$. Pour montrer que cette relation fait sens, il faut vérifier que la propriété $M \to N$ dans $\lambda_{\texttt{shift0}}$ implique $[\![M]\!] \leftrightarrow [\![N]\!]$ dans $(\!|\,\lambda\,|\!)$. $[\![M]\!] \leftrightarrow [\![N]\!]$ signifie que $[\![M]\!]$ peut être transformé en $[\![N]\!]$ par une série de réductions et d'expansions. Dans ce cas, la propriété de simulation implique que l'équivalence donnée par le calcul sous forme d'équation est préservée, c'est-à-dire $M = N$ dans $\lambda_{\texttt{shift0}}$ implique $[\![M]\!] = [\![N]\!]$ dans $(\!|\,\lambda\,|\!)$ (où $X = Y$ peut se lire comme $X \leftrightarrow Y$).

Il existe d'autres opérateurs de contrôle similaires à `shift0` et `reset0`. Un exemple souvent utilisé est celui de `shift` et `reset`. La différence peut se comprendre en observant les règles de réduction pour `shift0` and `shift`.

$$C[\texttt{reset0}\,(F[\texttt{shift0}\,V])] \to_{\texttt{shift0}} C[V\,(\lambda x.\,\texttt{reset0}\,(F[x]))]$$
$$C[\texttt{reset}\,(F[\texttt{shift}\,V])] \to_{\texttt{shift}} C[\texttt{reset}\,(V\,(\lambda x.\,\texttt{reset}\,(F[x])))]$$

`shift` conserve la délimitation de `reset` et installe une nouvelle délimitation dans la continuation. `shift0` est différent en ce qu'il élimine la délimitation de `reset0`. Dans tous les autres cas, la définition de $\lambda_{\texttt{shift}}$ (le λ-calcul pour l'appel par valeur construit avec `shift` et `reset`) est identique à celle de $\lambda_{\texttt{shift0}}$. La sémantique de `shift0` et `reset0` correspond elle à celle des *handlers* dans $(\!|\,\lambda\,|\!)$. Nous pouvons aussi transformer `shift` et `reset` dans $(\!|\,\lambda\,|\!)$, en transformant $\lambda_{\texttt{shift}}$ en $\lambda_{\texttt{shift0}}$.

L'interprétation de $[\![M]\!]_0$ un terme M de $\lambda_{\texttt{shift}}$ vers $\lambda_{\texttt{shift0}}$ est définie par :

$$[\![\texttt{reset}\,M]\!]_0 = \texttt{reset0}\,[\![M]\!]_0$$
$$[\![\texttt{shift}\,M]\!]_0 = \texttt{shift0}\,((\lambda m.\,\lambda x.\,\texttt{reset0}\,(m\,x))\,[\![M]\!]_0)$$
$$[\![M\,N]\!]_0 = [\![M]\!]_0\,[\![N]\!]_0$$
$$[\![\lambda x.\,M]\!]_0 = \lambda x.\,[\![M]\!]_0$$
$$[\![x]\!]_0 = x$$

Les plongements de $\lambda_{\texttt{shift0}}$ et $\lambda_{\texttt{shift}}$ vers $(\!|\,\lambda\,|\!)$ sont introduits. Dans ces deux transformations, les règles de réduction de $(\!|\,\lambda\,|\!)$ peuvent simuler celles de $\lambda_{\texttt{shift}}$ et $\lambda_{\texttt{shift0}}$. $(\!|\,\lambda\,|\!)$ n'est pas seulement défini sur les termes et les réductions, mais également sur un système de type. Le système de type de $(\!|\,\lambda\,|\!)$ peut être réutilisé pour $\lambda_{\texttt{shift}}$.

4.3 Deixis et quantification

Après cette longue présentation technique, il reste la question de l'utilisation de cette approche pour la modélisation de la sémantique de la langue.

Jirka Maršík propose dans (Maršík 2016 ; Maršík et Amblard 2016) de commencer avec un petit fragment contenant des noms propres et des verbes qui les prennent comme arguments.

$$\begin{aligned} \textsc{John}, \textsc{Mary} &: NP \\ \textsc{loves} &: NP \multimap NP \multimap S \end{aligned}$$

Et la sémantique qui leur est associée :

$$\begin{aligned} [\![\textsc{John}]\!] &= \eta\,\mathbf{j} \\ [\![\textsc{Mary}]\!] &= \eta\,\mathbf{m} \\ [\![\textsc{me}]\!] &= \texttt{speaker} \star (\lambda x.\, \eta\, x) \\ [\![\textsc{loves}]\!] &= \lambda OS.\, \mathbf{love} \cdot\!\gg S \ll\!\cdot\!\gg O \end{aligned}$$

Dans la sémantique de $[\![\textsc{me}]\!]$, L'opération `speaker` est introduite pour récupérer le locuteur courant et le rendre accessible pour la variable x. L'étoile ($\star$) utilisée avec `speaker` est une valeur par défaut du type *unit* 1.

$\cdot\!\gg$ et $\ll\!\cdot\!\gg$ sont des combinateurs qui simplifient les notations. De manière intuitive, ils sont construits à partir de $\gg\!=$. Il est possible de définir les schémas suivants :

$$\begin{aligned} &\_ \ll\!\cdot\ \_ : \mathcal{F}_E(\alpha \to \beta) \to \alpha \to \mathcal{F}_E(\beta) \\ &F \ll\!\cdot\ x = F \gg\!= (\lambda f.\, \eta\, (f\, x)) \\ &\_ \cdot\!\gg\ \_ : (\alpha \to \beta) \to \mathcal{F}_E(\alpha) \to \mathcal{F}_E(\beta) \\ &f \cdot\!\gg X = X \gg\!= (\lambda x.\, \eta\, (f\, x)) \\ &\_ \ll\!\cdot\!\gg\ \_ : \mathcal{F}_E(\alpha \to \beta) \to \mathcal{F}_E(\alpha) \to \mathcal{F}_E(\beta) \\ &F \ll\!\cdot\!\gg X = F \gg\!= (\lambda f.\, X \gg\!= (\lambda x.\, \eta\, (f\, x))) \end{aligned}$$

Toute la sémantique que nous voyons ici satisfait la propriété que si $M : \tau$, alors $[\![M]\!] : [\![\tau]\!]$. Soit, $[\![NP]\!] = \mathcal{F}_E(\iota)$ et $[\![S]\!] = \mathcal{F}_E(o)$, où ι et o sont des types d'individus et de propositions (respectivement).

Avec ces exemples, les phrases triviales suivantes peuvent être évaluées.

(4.1) John loves Mary.

(4.2) Mary loves me.

Dont les représentations sont :

$$[\![\textsc{loves}\ \textsc{Mary}\ \textsc{John}]\!] \twoheadrightarrow \eta\,(\mathbf{love}\,\mathbf{j}\,\mathbf{m}) \tag{4.3}$$

$$[\![\textsc{loves}\ \textsc{me}\ \textsc{Mary}]\!] \twoheadrightarrow \texttt{speaker} \star (\lambda x.\, \eta\,(\mathbf{love}\,\mathbf{m}\,x)) \tag{4.4}$$

La sémantique de 4.1, 4.3, est une proposition du type o incluse dans η, c'est-à-dire un élément qui peut être interprété dans un modèle. De la même manière, la sémantique

de 4.2, l'opérateur `speaker` est propagé à partir de l'entrée lexicale de ME jusqu'à la sémantique de toute la phrase. Ainsi, nous avons une expression algébrique qui a comme argument la proposition $\mathbf{love}\,\mathbf{m}\,x$ pour tous les $x : \iota$. Afin d'obtenir une seule proposition qui doit être considérée comme la valeur de vérité de la phrase et qui peut être évaluée dans un modèle, nous aurons besoin de définir le locuteur. Ceci est fait en introduisant un *handler* correspondant.

$$\begin{aligned}&\text{withSpeaker} : \iota \to \mathcal{F}_{\{\texttt{speaker}:1\rightarrowtail\iota\}\uplus E}(\alpha) \to \mathcal{F}_E(\alpha)\\&\text{withSpeaker} = \lambda sM.\, (\!|\, \texttt{speaker}\colon (\lambda xk.\, k\, s) \,|\!)\, M\end{aligned}$$

On notera que la clause η n'est pas reprise dans les $(\!|\ |\!)$ au dessus. Dans ce cas, nous supposons qu'il y a une interprétation par défaut : $\eta\colon (\lambda x.\, \eta\, x)$.

$$\text{withSpeaker}\ s\ [\![\text{LOVES ME MARY}]\!] \twoheadrightarrow \eta\,(\mathbf{love}\,\mathbf{m}\,s)$$

Jusqu'ici, une autre vision du problème aurait pu être d'introduire une constante qui représenterait le locuteur : **me**. Mais comme les *handlers* font partie de notre langage, il est possible de les inclure dans les entrées lexicales. Ils permettent donc de rattraper un phénomène, comme par exemple le discours direct (marqué par des guillemets), et qui permet de définir le locuteur à la volée.

$$\begin{aligned}\text{SAID}_{\text{IS}} &: S \multimap NP \multimap S\\ \text{SAID}_{\text{DS}} &: S \multimap NP \multimap S\end{aligned}$$

Ainsi, de nouveaux constructeurs syntaxiques sont disponibles : l'un pour l'utilisation de *said* en discours indirect et l'autre pour l'utilisation en discours direct. Leur sémantique est alors :

$$\begin{aligned}[\![\text{SAID}_{\text{IS}}]\!] &= \lambda CS.\, \mathbf{say} \cdot\!\gg S \ll\!\cdot\!\gg C\\ &= \lambda CS.\, S \gg\!= (\lambda s.\, \mathbf{say}\, s \cdot\!\gg C)\\ [\![\text{SAID}_{\text{DS}}]\!] &= \lambda CS.\, S \gg\!= (\lambda s.\, \mathbf{say}\, s \cdot\!\gg (\text{withSpeaker}\ s\, C))\end{aligned}$$

Nous avons élaboré les items lexicaux pour le discours indirect de sorte qu'il soit plus facile de les comparer avec ceux du discours direct. L'opérateur withSpeaker permet d'analyser :

(4.5) John said Mary loves me.

(4.6) John said, "Mary loves me".

$$[\![\text{SAID}_{\text{IS}}\ (\text{LOVES ME MARY})\ \text{JOHN}]\!] \twoheadrightarrow \texttt{speaker} \star (\lambda x.\, \eta\,(\mathbf{say}\,\mathbf{j}\,(\mathbf{love}\,\mathbf{m}\,x))) \tag{4.7}$$

$$[\![\text{SAID}_{\text{DS}}\ (\text{LOVES ME MARY})\ \text{JOHN}]\!] \twoheadrightarrow \eta\,(\mathbf{say}\,\mathbf{j}\,(\mathbf{love}\,\mathbf{m}\,\mathbf{j})) \tag{4.8}$$

La sémantique de la phrase 4.5, 4.7, dépend du locuteur (comme en témoigne l'utilisation de l'opérateur `speaker`) tandis que dans 4.8, la dépendance est limitée par l'utilisation du discours direct. On voit bien dans cet exemple comment les *handlers* permettent

de répartir les traitements entre les entrées lexicales d'une part, et également comment ils permettent de réaliser l'évaluation à différents moments du processus.

Jirka Maršík a par ailleurs proposé des opérateurs de portée de quantification QR dans le fragment. Le but est de permettre de projeter les effets d'un NP à l'extérieur de sa position habituelle. Il faut ajouter à la signature abstraite Σ^a_{QR} :

$$\mathrm{QR} : NP \to (NP \to S) \to S$$

Puis nous donnons une sémantique à l'opérateur QR dans $\mathcal{L}_{QR}$:

$$[\![\mathrm{QR}]\!] := \lambda Xk.\ \mathrm{SI}\ (X \gg\!= (\lambda x.\ k\ (\eta\ x)))$$

SI est utilisé pour *Scope Island* puisque justement le *handler* va rattraper le calcul de la portée pour la figer. Cela nous permet de calculer l'autre portée des quantificateurs pour l'exemple 4.9.

(4.9) *Every man loves a woman* Tout homme aime une femme

$$\begin{aligned}
&[\![\mathrm{QR}\ (\textsc{a woman})\ (\lambda x.\ \textsc{loves}\ x\ (\textsc{every man}))]\!] \\
\to^* &\ \mathrm{SI}\ (SCOPE\ (\lambda k.\ \overline{\exists}\, x.\ (\eta\ (\mathbf{woman}\ x))\ \overline{\wedge}\ (k\ x))\ (\lambda x.\ \eta\ (\forall y.\ \mathbf{man}\ y \to \mathbf{love}\ y\ x))) \\
\to^* &\ (\lambda k.\ \overline{\exists}\, x.\ (\eta\ (\mathbf{woman}\ x))\ \overline{\wedge}\ (k\ x))\ (\lambda x.\ \eta\ (\forall y.\ \mathbf{man}\ y \to \mathbf{love}\ y\ x)) \\
\to^* &\ \eta\ (\exists x.\ \mathbf{woman}\ x \wedge (\forall y.\ \mathbf{man}\ y \to \mathbf{love}\ y\ x))
\end{aligned}$$

L'opérateur QR qui apparaît dans la dénotation de la phrase nous permet de déplacer la portée d'un quantificateur introduit par un NP, (Farkas 1981).

Une autre solution est de permettre à un quantificateur de s'échapper de manière non déterministe. Cette solution est moins directe et elle contrôle moins finement la dérivation, mais elle est plus simple à déployer. Pour cela nous introduisons un nouvel effet : $CHOOSE : 1 \to 2 \in E$, où 2 est le type des booléens (c'est à dire qui peuvent être soit vrai soit faux : $1 + 1$), ainsi que la forme syntaxique suivante pour les calculs non-déterministes :

$$\begin{aligned}
A + B = CHOOSE\ *\ (\lambda b.\ \mathrm{case}\ b\ \mathrm{of}\ &\ \mathrm{inl}(*) \Rightarrow A \\
&\mid\ \mathrm{inr}(*) \Rightarrow B)
\end{aligned}$$

où inl et inr sont deux injections, l'une vers la composante gauche et l'autre droite de l'opérateur $+$.

Cet exemple est en fait réel, car certains quantificateurs (ceux des indéfinis) doivent pouvoir s'échapper des *scopes island*. Une variante de $SCOPE$ (en tant qu'effet) est introduite :

$$SCOPE_P : ((\iota \to \mathcal{F}(o)) \to \mathcal{F}(o)) \to \iota \in E$$

Nous définissons l'extension $E_{Q'}$, construite à partir de E_Q :

$$E_{Q'}(\langle \Sigma^a, \Sigma^o, \mathcal{L} \rangle) = \langle \Sigma^a, \Sigma^o, \mathcal{L} \uplus \mathcal{L}_{Q'} \rangle$$

où les changements dans le lexique, $\mathcal{L}_{Q'}$, sont les suivants :

$$\begin{aligned}
[\![\text{EVERY}]\!] &:= \lambda N.\ SCOPE\ (\lambda k.\ \text{SL}\ (\overline{\forall}\ (\lambda x.\ (N \ll\cdot\ x) \Rightarrow (k\ x))))\ \eta \\
[\![\text{SOME}]\!] := [\![\text{A}]\!] &:= \lambda N.\ SCOPE_P\ (\lambda k.\ \text{SL}\ (\overline{\exists}\ (\lambda x.\ (N \ll\cdot\ x) \overline{\wedge} (k\ x))))\ \eta \\
\text{SI} &= \mathcal{L}(\text{SI}) \circ \text{SL} \\
\text{SL} &: \mathcal{F}(o) \to \mathcal{F}(o) \\
\text{SL} &= [\mathcal{H}\ (SCOPE_P\ (\lambda ck.\ (c\ k) + (SCOPE_P\ c\ k)))]
\end{aligned}$$

On constate qu'il n'est pas nécessaire de modifier de nombreuses entrées. Le *handler* SL (*Scope Location*) introduit une position pour qu'un quantificateur prenne sa portée ou s'échappe.

(4.10) Every man loves a woman

Pour aborder l'exemple 4.10, les notations suivantes sont nécessaires :

$$\begin{aligned}
c_1 &= \lambda k.\overline{\forall}\ (\lambda x.\ (\eta\ (\mathbf{man}\ x)) \Rightarrow (k\ x)) \\
c_2 &= \lambda k.\ \overline{\exists}\ (\lambda x.\ (\eta\ (\mathbf{woman}\ x)) \overline{\wedge} (k\ x)) \\
[\![\text{EVERY MAN}]\!] &= SCOPE\ (\text{SL} \circ c_1)\ \eta \\
[\![\text{A WOMAN}]\!] &= SCOPE_P\ (\text{SL} \circ c_2)\ \eta
\end{aligned}$$

Ainsi, la possibilité pour un quantificateur de reprendre sa portée apparaît à la fin

de la dérivation suivante :

$$
\begin{aligned}
&[\![\textsc{loves (a woman) (every man)}]\!] \\
\to^* &\ \mathrm{SI}\ (\mathit{SCOPE}\ (\mathrm{SL} \circ c_1)\ (\lambda x.\ \mathit{SCOPE}_P\ (\mathrm{SL} \circ c_2)\ (\lambda y.\ \eta\ (\mathbf{love}\ x\ y)))) \\
\to^* &\ \mathrm{SL}\ (c_1\ (\lambda x.\ \mathrm{SI}\ (\mathit{SCOPE}_P\ (\mathrm{SL} \circ c_2)\ (\lambda y.\ \eta\ (\mathbf{love}\ x\ y))))) \\
\to^* &\ \mathrm{SL}\ (c_1\ (\lambda x.\ \mathrm{SL}\ (c_2\ (\lambda y.\ \mathrm{SI}\ (\eta\ (\mathbf{love}\ x\ y)))) + \mathit{SCOPE}_P\ (\mathrm{SL} \circ c_2)\ (\lambda y.\ \mathrm{SI}\ (\eta\ (\mathbf{love}\ x\ y))))) \\
\to^* &\ \mathrm{SL}\ (c_1\ (\lambda x.\ \mathrm{SL}\ (c_2\ (\lambda y.\ \eta\ (\mathbf{love}\ x\ y))) + \mathit{SCOPE}_P\ (\mathrm{SL} \circ c_2)\ (\lambda y.\ \eta\ (\mathbf{love}\ x\ y)))) \\
\to^* &\ \mathrm{SL}\ (c_1\ (\lambda x.\ \mathrm{SL}\ (\eta\ (\exists y.\ \mathbf{woman}\ y \wedge \mathbf{love}\ x\ y) + \mathit{SCOPE}_P\ (\mathrm{SL} \circ c_2)\ (\lambda y.\ \eta\ (\mathbf{love}\ x\ y)))) \\
\to^* &\ \mathrm{SL}\ (c_1\ (\lambda x.\ (\eta\ (\exists y.\ \mathbf{woman}\ y \wedge \mathbf{love}\ x\ y) + \mathit{SCOPE}_P\ (\mathrm{SL} \circ c_2)\ (\lambda y.\ \eta\ (\mathbf{love}\ x\ y)))) \\
\to^* &\ \mathrm{SL}\ (\eta\ (\forall x.\ \mathbf{man}\ x \to (\exists y.\ \mathbf{woman}\ y \wedge \mathbf{love}\ x\ y)) \\
&\qquad + \mathit{SCOPE}_P\ (\mathrm{SL} \circ c_2)\ (\lambda y.\ \eta\ (\forall x.\ \mathbf{man}\ x \to \mathbf{love}\ x\ y))) \\
\to^* &\ \mathrm{SL}\ (\eta\ (\forall x.\ \mathbf{man}\ x \to (\exists y.\ \mathbf{woman}\ y \wedge \mathbf{love}\ x\ y))) \\
&+ \mathrm{SL}\ (\mathit{SCOPE}_P\ (\mathrm{SL} \circ c_2)\ (\lambda y.\ \eta\ (\forall x.\ \mathbf{man}\ x \to \mathbf{love}\ x\ y))) \\
\to^* &\ \eta\ (\forall x.\ \mathbf{man}\ x \to (\exists y.\ \mathbf{woman}\ y \wedge \mathbf{love}\ x\ y)) \\
&+ \mathrm{SL}\ (c_2\ (\lambda y.\ \eta\ (\forall x.\ \mathbf{man}\ x \to \mathbf{love}\ x\ y))) \\
&+ \mathit{SCOPE}_P\ (\mathrm{SL} \circ c_2)\ (\lambda y.\ \eta\ (\forall x.\ \mathbf{man}\ x \to \mathbf{love}\ x\ y))) \\
\to^* &\ \eta\ (\forall x.\ \mathbf{man}\ x \to (\exists y.\ \mathbf{woman}\ y \wedge \mathbf{love}\ x\ y)) \\
&+ \mathrm{SL}\ (\eta\ (\exists y.\ \mathbf{woman}\ y \wedge (\forall x.\ \mathbf{man}\ x \to \mathbf{love}\ x\ y))) \\
&+ \mathit{SCOPE}_P\ (\mathrm{SL} \circ c_2)\ (\lambda y.\ \eta\ (\forall x.\ \mathbf{man}\ x \to \mathbf{love}\ x\ y)) \\
\to^* &\ \eta\ (\forall x.\ \mathbf{man}\ x \to (\exists y.\ \mathbf{woman}\ y \wedge \mathbf{love}\ x\ y)) \\
&+ \eta\ (\exists y.\ \mathbf{woman}\ y \wedge (\forall x.\ \mathbf{man}\ x \to \mathbf{love}\ x\ y)) \\
&+ \mathit{SCOPE}_P\ (\mathrm{SL} \circ c_2)\ (\lambda y.\ \eta\ (\forall x.\ \mathbf{man}\ x \to \mathbf{love}\ x\ y))
\end{aligned}
$$

Cette dérivation un peu longue est obtenue par β-réduction et application des règles du calcul. Il est alors possible d'introduire le traitement d'un phénomène particulier en répartissant sur différents termes son traitement.

Cette présentation est complexe car le cadre est riche en propriétés. Dans (Maršík 2016), Jirka Maršík a construit étape par étape le traitement de la deixis et des quantificateurs que nous venons d'aborder, mais également les implicatures conventionnelles (ce qui est signifié implicitement par un locuteur), l'accessibilité des référents de discours pour la résolution des anaphores pronominales et la présupposition. Pour chacun de ces traitements il s'est astreint à définir une méthodologie qui utilise toute la flexibilité du formalisme. Par ailleurs, il a montré comment ces traitements pouvaient mécaniquement être rassemblés dans une unique grammaire sémantique.

Deuxième partie

Discours et interprétation

Chapitre 5

Interprétation de la sémantique : le projet SLAM

Sommaire

Avec ce nouveau chapitre nous débutons une nouvelle partie de la présentation de notre travail de recherche. Il nous est apparu nécessaire de changer de perspective quant à notre vision des modèles formels en les confrontant à des cas concrets. L'opportunité d'une collaboration interdisciplinaire sur l'analyse de la production langagière de schizophrènes s'est présentée. Cet aspect nous est apparu pertinent car s'il permet d'interroger l'adéquation des propositions théoriques à l'expression d'une réalité, il permet aussi d'étudier le sens des modèles dans une perspective cognitive (et donc au delà de leur cohérence théorique). Nous avons travaillé dans cette perspective, tout en constatant que s'il était possible d'utiliser les aspects théoriques, un important travail s'ouvrait pour trouver des liens effectifs entre troubles du langage et troubles de la pensée.

Un aspect de nos recherches concerne la mise en œuvre de la modélisation formelle du discours. Des collaborations avec Michel Musiol (psychologue) et Manuel Rebuschi (philosophe) ont été initiées dans le cadre de l'opération de recherche de la MSH-Lorraine DIARAFOR - DIAlogues, RAtionalités et FORmalismes - Etudes croisées logique / psychologie / épistémologie - pour la modélisation formelle de dialogues avec des patients souffrant de troubles de la pensée. Ces échanges sont porteurs de dysfonctionnements. Le choix de la pathologie est fortement lié aux opportunités scientifiques. La proposition permettait de répondre à la motivation de l'étude d'un terrain empirique qui, en plus, est rapidement apparu pertinent pour nos recherches.

Il s'est révélé, et nous allons revenir longuement sur ces aspects dans les prochains chapitres, que les schizophrènes réalisent une perturbation du langage manifeste d'un dysfonctionnement de la pensée. Sur ces problèmes, nous pouvons formuler des hypothèses sur les processus en œuvre. Si nous pouvons expliquer pourquoi et comment les troubles sont liés, alors nous pouvons avancer des arguments sur le fonctionnement cognitif. En général, l'étude du dysfonctionnement nous renseigne sur le fonctionnement normal. Ce chapitre explicite le contexte du projet et la constitution de la ressource à la base de nos travaux. Le chapitre 6 reprend les positions théoriques de notre analyse. Le chapitre 7 présente un outil que nous développons pour l'analyse quantitative de la ressource.

5.1 Contexte

Les discussions dans le cadre de DIARAFOR ont rapidement donné lieu à des échanges constructifs qui se sont concrétisés dans le projet « Schizophrénie et Langage : Analyse et Modélisation » (SLAM). Ces travaux ont été hébergés à la MSH-Lorraine de 2012 à 2015 et ont reçu le soutien du CNRS sous la forme d'un PEPS [1] en 2013 et 2014. Le projet a été par la suite co-financé par le CPER [2] Lorraine, l'Université de Lorraine et la région Grand-Est. Grâce à ces soutiens, nous avons pu activement travailler et échanger avec d'autres collègues internationaux, en particulier dans le cadre de la conférence « (In)Coherence du discours » que nous avons organisée trois fois depuis 2013 et dont la quatrième édition est annoncée. Nous nous sommes également réunis pendant le workshop organisé en l'honneur de Hans Kamp (Octobre 2014).

Le projet SLAM vise à systématiser l'étude et la formalisation des conversations pathologiques entre des patients schizophrènes et des psychologues, dans le cadre d'une approche interdisciplinaire (psychologie, linguistique, informatique). La figure 5.1 en présente les grandes orientations. Le travail est organisé selon plusieurs axes, sur lesquels nous reviendrons :

- constituer et gérer une ressource linguistique sur la pathologie mentale constituée des transcriptions [3] des entretiens. Nous travaillons sur des corpus spécifiques et nous avons cherché à développer des outils numériques pour la collecte (dans la perspective de pouvoir rendre public un tel corpus, ce que nous n'avons finalement pas pu réaliser) ;
- conduire des études épistémologiques et philosophiques sur les concepts de norme, de folie ou de rationalité ;
- identifier les usages pathologiques (dysfonctionnements linguistiques) par l'utilisation des théories et outils du traitement automatique des langues.

Le projet s'est ainsi naturellement structuré en trois composantes autour de chacune de ces problématiques. Dans un premier temps nous posons le contexte de cette étude dans la section 5.1, et nous exposons la constitution de la ressource dans la section 5.2. Puis nous revenons en détail sur les axes sur lesquels nous nous sommes plus particulièrement investis. En particulier, le chapitre 6 introduit les enjeux de la modélisation formelle et le chapitre 7 le travail d'étude linguistique sur les propriétés du corpus, notamment les analyses quantitatives de ces données. Les aspects épistémologiques et philosophiques apparaissent de manière transverse dans nos publications. Notre contribution plus spécifique reste modeste sur ces aspects.

5.1.1 Des discontinuités décisives à l'aide au diagnostic

1. Projet Exploratoire Premier Soutien de la Mission pour l'Interdisciplinarité dans le cadre du programme HuMaIn - Humanité Mathématiques Informatique.

2. Contrat Plan État Région.

3. Comme nous l'avons mentionné précédemment, la transcription en linguistique est l'opération qui consiste à substituer à chaque son d'une langue un graphème. Il s'agit ici de produire une version écrite des enregistrements des entretiens.

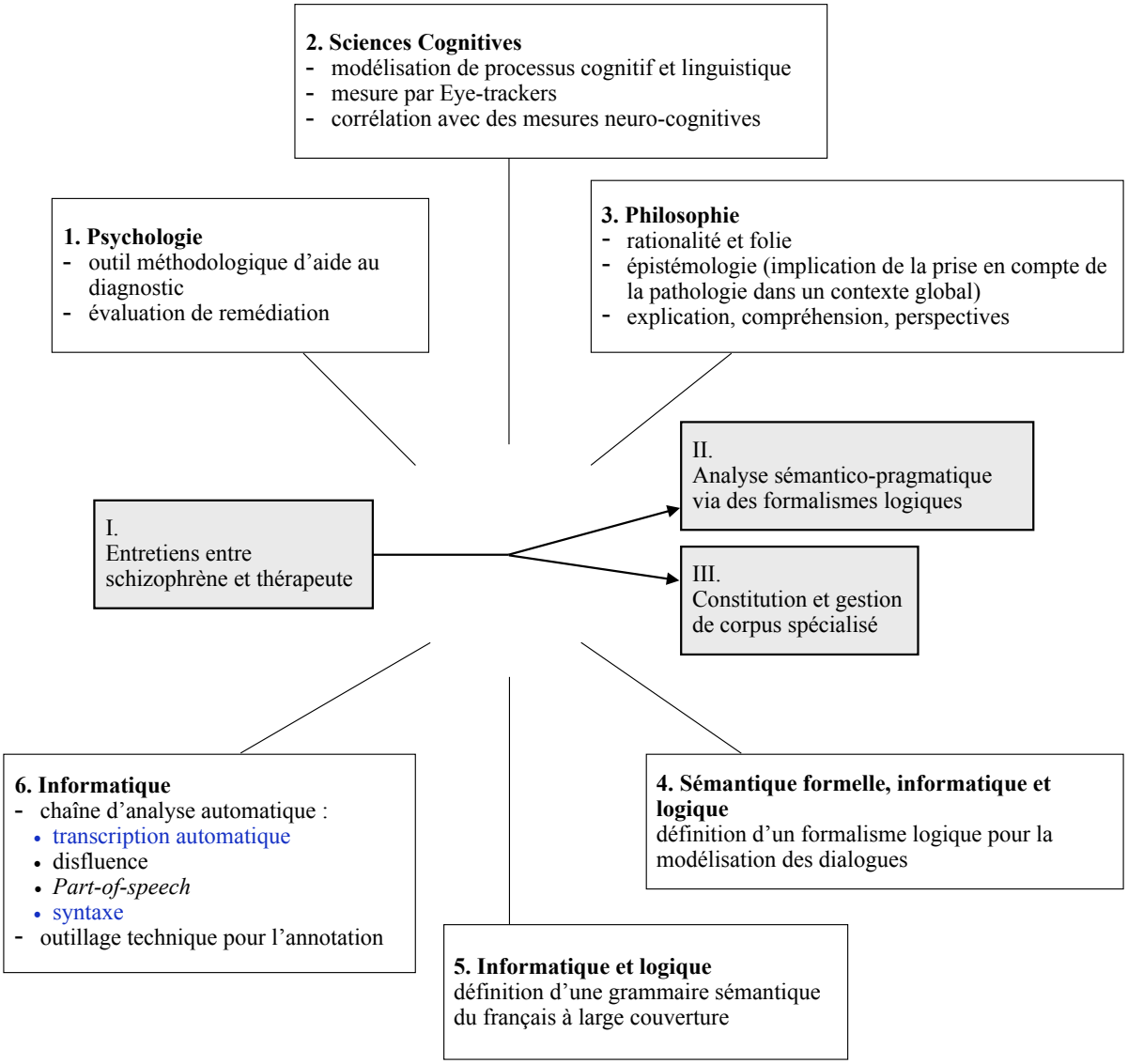


FIGURE 5.1 – Schéma général du projet SLAM

L'objet du projet SLAM concerne par nature des problématiques sociétales. Les résultats produits peuvent être considérés comme la mise en évidence d'indices objectifs dans la manifestation de la pathologie mentale. La méthodologie que nous avons définie peut être transférée vers les praticiens psychiatres pour les aider à argumenter un diagnostic. Il ne s'agit en aucun cas de considérer que l'outil va remplacer le spécialiste mais bien venir en soutien de son expertise. Une autre utilisation réside dans l'accompagnement d'un patient dans un parcours de soins. Il s'agit de modéliser ces propositions pour mesurer l'avancée de sa remédiation, voire proposer des exercices pour cette remédiation.

Le fait de travailler sur du matériel linguistique produit par des patients impose des contraintes fortes pour la gestion de la ressource. Proposer une analyse formelle de la langue implique d'introduire une vision normée de l'usage de la langue. Une déviation du comportement serait alors une manifestation d'un dysfonctionnement et peut être interprétée comme un symptôme. Or, tout locuteur est confronté quotidiennement à des troubles du language sans qu'ils ne proviennent de troubles de la pensée. La pose d'un diagnostic peut avoir un impact significatif sur la vie des patients, elle ne peut donc pas souffrir d'incertitudes.

Nous nous intéressons à la production langagière des schizophrènes car elle fait apparaître un dysfonctionnement à un niveau abstrait de description linguistique. Le schizophrène semble s'appuyer sur un élément porteur d'une ambiguïté linguistique, par exemple un terme polysémique, et après que cette ambiguïté ait été résolue, il le reprend en choisissant une autre interprétation. Ainsi, la chaîne de coréférences qui permet de former un discours cohérent n'est pas respectée et l'autre interlocuteur ne comprend plus l'interaction. Si certaines de ces discontinuités sont parfois pragmatiquement réparables, d'autres sont trop complexes et bloquent la compréhension. On les appelle « discontinuités décisives ». L'ensemble des analyses que nous avons portées sur la première partie du corpus est présenté dans (Rebuschi, Amblard et Musiol 2012).

G82 l'an dernier euh (→) j'savais pas comment faire **j'étais perdue**$_1$ et pourtant j'avais pris mes médicaments j'suis dans un état vous voyez même ma bouche elle est sèche j'suis dans un triste état
V83 Vous êtes quand même bien (↑)
G84 J'pense que ma tête est bien mais on croirait à moitié (↓) la moitié qui va et la moitié qui va pas j'ai l'impression de ça vous voyez (↑)
V85 D'accord
G86 Ou alors c'est la conscience peut-être la conscience est ce que c'est ça (↑)
V87 Vous savez **ça arrive à tout le monde d'avoir des moments biens et des moments où on est perdu**$_2$
G88 **Oui j'ai peur de perdre**$_3$ **tout le monde**

FIGURE 5.2 – Exemple de discontinuité décisive

La figure 5.2 reprend l'un de ces exemples, en présentant un extrait de la transcription d'un entretien contenant une de ces discontinuités. Ici, le terme pivot est la lexie **perdre**, d'abord utilisée pour **être perdu**$_1$ qui glisse à l'interprétation **perdre**$_3$ **tout**

le monde. Le premier usage de **perdre** est cohérent avec le deuxième, le deuxième est interprétable par rapport au troisième, mais le premier n'a plus de sens par rapport au troisième. Si on assume que la cohérence doit être valide tout au long des usages du terme, l'interprétation doit être transitive. Or justement ici, nous n'avons pas la transitivité entre les interprétations, ce qui rend la compréhension impossible. Cette propriété permet de définir l'un des critères d'identification des ruptures dites décisives.

Les autres exemples montrent que le problème peut être entendu à des niveaux linguistiques très différents partant de la phonologie (Pro par Vocation *vs* provocation) à la pragmatique (modification du référent de discours sur un nom propre).

5.1.2 Le projet des différentes annotations

L'objet du projet est de proposer des analyses de ces entretiens tant au niveau sémantique que pragmatique, et de s'intéresser à la qualité linguistique de la production. Il s'agit donc de produire une version de la ressource annotée sur plusieurs niveaux linguistiques.

Une première phase du travail a consisté en l'étude des différents scénarii méthodologiques. Trois éléments qui structurent l'annotation sont apparus :

— la nécessité de disposer d'une transcription de très bonne qualité ;
— la possibilité d'utiliser des outils automatiques pour annoter certains niveaux ;
— la nécessité de travailler au système d'annotation et à son outillage pour le niveau sémantico-pragmatique.

La figure 5.3 donne une représentation de la position de ces éléments les uns par rapport aux autres. La transcription convertit le son en une version textuelle. Pour cela nous avons travaillé à définir spécifiquement un guide d'annotations, inspiré de (Blanche-Benveniste 1997) qui normalise les transcriptions produites grâce au logiciel `Transcriber`, (Barras et al. 1998). Cette étape est considérée comme la première étape d'annotations de la ressource. La transcription est un point fondamental pour le reste de la recherche, sa qualité doit être la meilleure possible, c'est pourquoi nous utilisons une transcription manuelle. Par ailleurs, nous avons procédé à des tests qualitatifs en faisant transcrire un même extrait par plusieurs transcripteurs pour calculer des accords inter-annotateurs.

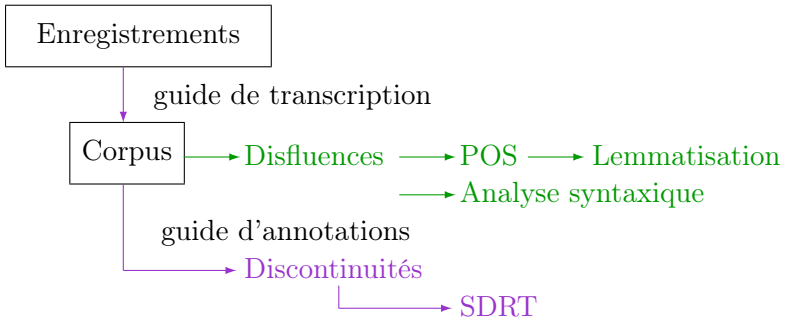


FIGURE 5.3 – Organisation des niveaux d'analyse de la ressource SLAM

Comme nous l'avons déjà indiqué, nous avons identifié dans ces transcriptions des enchaînements de tours de parole dont l'interprétation n'était plus possible et qui formaient des « ruptures décisives », (Musiol et Trognon 1996). Nous nous sommes interrogés sur les méthodologies à déployer pour analyser le plus largement possible ces données et ne pas les aborder uniquement du point de vue sémantico-pragmatique. Nous avons donc décidé d'utiliser des outils du traitement automatique des langues pour produire d'autres annotations. Il s'agit des éléments en vert de la figure 5.3. Dans un premier temps, il nous est paru nécessaire d'analyser les marques spécifiques de l'oral que sont les phénomènes de disfluences (réalisations orales qui rompent la continuité syntaxique). Quatre formes sont prédominantes dans les corpus oraux : les interjections d'hésitation (*euh*), les répétitions, les autocorrections immédiates et les amorces de morphèmes. Dans le même temps, nous avons écarté de notre étude les aspects explicitement phonologiques et phonétiques, non pas par manque d'intérêt mais de disponibilité de compétences. Ces questions seraient des pistes de recherche tout aussi pertinentes que celles que nous avons décidé de poursuivre.

Une fois ces informations obtenues, nous nous sommes concentrés sur les étiquetages morpho-syntaxiques (*part-of-speech (POS)*) et des lemmes. Un étiquetage morpho-syntaxique est un processus qui associe à chaque mot d'un énoncé une information grammaticale sur sa fonction. Il est aussi possible d'y associer des éléments morphologiques comme des propriétés de genre, nombre, *etc.*. De manière simplifiée, le lemme associé à un mot d'un énoncé est la forme dont il est dérivé, par exemple le lemme de « relieront » est le verbe « relier ». Nous envisageons de poursuivre le travail avec d'autres éléments, en particulier des analyses syntaxiques.

Pour certains autres niveaux, les outils ne proposent pas une qualité suffisante. Ainsi nous avons décidé que certains niveaux d'analyse seraient abordés par des annotations manuelles, c'est-à-dire des annotations produites par des humains. Le travail étant particulièrement couteux, nous nous sommes limités à travailler et outiller les annotations pour identifier des passages contenant des discontinuités décisives et, pour ces passages, d'en produire des représentations inspirées de la SDRT. Ces éléments apparaissent en violet sur la figure 5.3.

Ces différents niveaux d'annotations nous interrogent de manière très différentes. Les attendus sur la précision et la qualité des annotations ne sont pas du tout les mêmes, d'autant que les volumes de données sont très différents. Par exemple, les annotations en disfluences ou en catégories morpho-syntaxiques produisent beaucoup d'informations. La précision et le rappel [4] peuvent être inférieurs par rapport à ceux du niveau sémantico-pragmatique qui doivent être de très bonne qualité.

4. La précision et le rappel sont deux mesures classiques permettant d'évaluer la qualité d'un outil automatique. Il s'agit de mesurer si tous les éléments sont reconnus, et si ceux reconnus le sont correctement.

5.2 La ressource

Il est difficile d'expliciter la constitution de la ressource comme s'il s'agissait d'un objet unique. Dans les différentes publications relatives à cet aspect de nos recherches, le terme de ressource recouvre des réalités différentes. Nous revenons dans cette section sur les difficultés inhérentes à sa constitution et nous présentons les différentes parties qui la composent actuellement.

5.2.1 Le protocole méthodologique

Le cœur du protocole méthodologique qui nous concerne est l'enregistrement d'un entretien semi-dirigé entre un patient schizophrène et un psychologue. Ce type d'entretien a ceci de particulier que le psychologue n'est pas personnellement engagé dans l'échange, sa mission est de maintenir le dialogue en produisant des relances. Par contre, le patient est là pour revenir sur différents éléments biographiques ou sur ses conditions actuelles de vie. Le psychologue a pour fonction d'aider le patient à verbaliser une description de sa situation. On considère alors qu'il s'agit d'une forme de discours (séquences d'énoncés qui forment un tout cohérent) qui doit respecter les éléments traditionnels de la structure des discours.

Pour avoir la possibilité de rencontrer des patients schizophrènes, il est nécessaire de passer par les institutions médicales qui ont posé le diagnostic. Or, pour entrer dans le milieu médical et accéder aux patients, il est nécessaire de disposer de l'autorisation de la Commission de la Protection de la Personne (CPP). Il y est vérifié la nature des recherches conduites, leur aspect intrusif, tant du point de vue personnel que physique, ainsi que la nécessité de disposer des données pour la recherche. Une fois que ces autorisations sont accordées, le protocole décrit ne peut plus être modifié. Par ailleurs, ces autorisations impliquent de disposer d'assurances particulières ainsi que de déposer des demandes d'autorisation auprès de la CNIL [5].

Avant de pouvoir disposer des entretiens, il a donc fallu avoir obtenu ces autorisations, avoir arrêté le protocole, contracté des assurances et enfin envoyé des psychologues dans les hôpitaux partenaires. Ces éléments ont une force d'inertie non négligeable qui retarde considérablement l'avancée de la recherche et qui explique les délais pour la constitution de la ressource. Ce système reste absolument nécessaire pour protéger les patients.

Étant données les contraintes pour obtenir une autorisation, il n'apparaît pas raisonnable d'en demander une pour simplement enregistrer les entretiens. Nous nous sommes donc intégrés dans un protocole plus large qui consistait entre autres à réaliser des tests neuro-cognitifs pour disposer de mesures sur les capacités cognitives des patients. Les tests neuro-cognitifs utilisés sont :

1. *Wechsler Adult Intelligence Scale-III* (mesure du quotient intellectuel, ou QI) (Wechsler 1958) ;
2. *California Verbal Learning Test* (capacités cognitives et de stratégie) (Woods et al. 2006) ;

5. Commission Nationale Informatique et Liberté https://www.cnil.fr/

	La Rochelle	Le Vinatier, Lyon	Hopital Saint-Antoine, Paris	Aix-en-Provence
Enregistrement de l'entretien	x	x	x	x
Tests neuro-cognitifs		x	x	x
Avec double eye-trackers		x	x	x
Second entretien avec EEG			x	

TABLE 5.1 – Présentation des différentes parties de la ressource en fonction des lieux de recueil des données, présentées dans l'ordre chronologique

3. *Trail Making Test* (dépréciation de la flexibilité cognitive et de l'inhibition, déficit qui peut affecter la vitesse du système perceptif moteur, la flexibilité spontanée ou la flexibilité de réaction) (Bowie et Harvey 2006).

Nous avons également décidé d'accompagner l'enregistrement de l'entretien de la mesure des mouvements des yeux avec un oculomètre (*eye-trackers*). Cet appareillage a été utilisé pour tester une hypothèse ancienne de la littérature selon laquelle les schizophrènes ont un comportement oculo-moteur spécifique (Lindsey et al. 1978 ; Stark 1983 ; Levy, Gooding et O'Driscolln 2010). Il s'agissait de valider cette hypothèse, tout en vérifiant que les déclencheurs de ces saccades oculaires n'étaient pas dues au regard du psychologue. Pour mesurer cela, le psychologue est lui aussi enregistré par un oculomètre. Le résultat a été repris dans (Padrovni 2015) dans le cadre de sa thèse sous la direction de Michel Musiol par l'exploitation d'une partie de ces données.

Par ailleurs, dans une seconde vague de constitution de corpus, nous avons souhaité tester des hypothèses relatives à l'activité du cerveau en utilisant un système qui mesure l'activité électrique de l'encéphalogramme (électroencéphalographie - EEG). Ainsi, en plus des tests neuro-cognitifs d'un premier entretien avec le double système d'eye-trackers, le patient devait réaliser un second entretien avec un EEG. On peut noter dès à présent que le protocole ainsi obtenu est particulièrement complexe et lourd à mettre en œuvre. Nous revenons sur les difficultés relatives à la constitution d'une telle ressource dans la section suivante.

Nous avons dès lors pu lancer plusieurs campagnes de récupération des données. Nous reprenons dans la table 5.2 la répartition des patients et témoins dans les différentes campagnes, et dans la table 5.3 le volume de données pertinentes récupérées. À chaque fois, nous avons utilisé un partenaire hospitalier spécifique qui acceptait de participer à l'étude. Dans le tableau 5.1, les différents lieux sont associés aux différents recueils de données effectués. Les trois premiers ont déjà été réalisés et le dernier est l'actuel projet de recueil. À la lecture de ce tableau, on constate que nous sommes allés vers une

complexité croissante du protocole en ajoutant des éléments à tester, et que nous nous dirigeons maintenant vers une simplification en en supprimant.

5.2.2 Difficultés de constitution de la ressource

Comme nous l'avons mentionné, ce travail de constitution de la ressource s'est avéré plus complexe qu'attendu. Au delà des aspects pratiques et administratifs comme l'obtention des autorisations et du fait de disposer des ressources financières pour contracter les assurances et envoyer les personnels sur place, il ne faut pas négliger l'importance du protocole. Comme le montre le tableau 5.1, nous faisons évoluer notre protocole à chaque étape de recueil des données. La tradition propriétaire des ressources dans ces contextes explique cette évolution dynamique. En effet, la communauté académique en psychologie n'a pas l'habitude de rendre publiques ses ressources, ni, bien souvent, tous les éléments de ses protocoles. Il est donc très difficile de pouvoir se comparer à des études antérieures pour situer ses recherches tant du point de vue méthodologique que qualitatif.

La réussite de notre recueil dépend aussi de l'investissement des partenaires hospitaliers qui nous donnent accès aux patients. Ils doivent identifier et adresser aux psychologues du projet les patients correspondant à l'étude, il faut parallèlement que le psychologue et le patient soient physiquement présents et disponibles en même temps. Il n'est pas toujours possible de laisser un psychologue à temps complet dans un service, tant pour le coût induit que pour éviter de désorganiser les services. Beaucoup d'entretiens n'ont de ce fait pas pu être réalisés. Il est aussi fréquent que des patients acceptent de participer, mais finalement ne se présentent pas à leur rendez-vous, ce qui peut s'expliquer compte tenu de la pathologie étudiée.

Pour la phase de recueil à Lyon, nous avons mesuré que le taux de refus de participation à l'étude était de 45% (18). Cette quantité peut paraître importante, mais il est souvent difficile de convaincre les patients de participer, sachant qu'ils n'ont aucun bénéfice direct à leur participation. Cet aspect est primordial car il garantit une certaine honnêteté dans leur implication et donc une véracité des propos. Nous sommes attachés à ce que l'entretien puisse se dérouler sans contrainte, ainsi rien de ce qui est échangé ne peut être utilisé pour ou contre les patients (dans la limite de la déontologie du praticien). Le taux d'abandon des patients de la cohorte pour l'étude était de 10% (4). Ces patients correspondent principalement à ceux ne se présentant pas pour l'ensemble du protocole. Finalement, 45% (18) patients ont participé intégralement à l'étude. Si on observe la quantité de patients mobilisés dans cette phase, ces taux sont très bons. Ils sont principalement dus à trois éléments :

- la disponibilité sur des temps longs de la psychologue, qui pouvait également participer à d'autres activités du service et donc se faire connaître des patients ;
- le fait que ces patients étaient dans un service de remédiation, et donc plutôt dans des états physiques et psychologiques leur permettant de participer à l'étude ;
- la qualité de l'investissement de l'encadrement hospitalier.

À titre de comparaison, la dernière phase de recueil effective à Paris n'a permis de réaliser que cinq protocoles pour l'étude. Ce nombre est très faible, mais il s'explique par plusieurs circonstances dans le déroulement du projet (dont la faible disponibilité des

	La Rochelle			Le Vinatier, Lyon			Hopital Saint-Antoine, Paris			Aix-en-Provence			Total
	♂	♀	tot	♂	♀	tot	♂	♀	tot	♂	♀	tot	
Schizophrènes													
sous traitement	15	3	18	21	3	24	3	2	5				
sans traitement				1	6	7							
total			18			31			5				54
Témoins	15	8	23	4	4	8							31
Total	30	11	41	26	13	39			5				85

TABLE 5.2 – Nombre d'entretiens par phase de recueil des corpus

patients en ambulatoire). L'argument principal est certainement que le protocole était particulièrement lourd à mettre en œuvre, ce qui a découragé plusieurs candidats. Il faut relativiser ce faible nombre en considérant que pour chaque patient deux entretiens ont été réalisés. Étant donné l'échec relatif de cette phase, nous n'avons pas lancé le recueil de données pour les témoins, qui seront inclus dans la phase suivante.

Nous avons écarté les données recueillies à Paris pour poursuivre notre travail et nous avons utilisé les enregistrements des deux premiers recueils. Par ailleurs nous nous sommes concentrés uniquement sur les enregistrements (et par la suite leur transcription), écartant pour le moment les autres données récupérées ou les laissant à l'analyse d'autres collègues. La phase à Aix-en-Provence est déjà constituée de sept entretiens.

Nous travaillons ainsi sur un corpus de transcription de 49 entretiens (54 moins les 5 patients de Paris), ce qui, dans un tel contexte n'est pas négligeable. Par ailleurs, le corpus transcrit contient environ 375 000 mots. Nous avons procédé à une présentation fine de ce corpus dans (Amblard, Fort, Demily et al. 2015). Le chapitre suivant reprend la problématique de la modélisation formelle.

	Corpus La Rochelle				Corpus Lyon			
	nb tours		nb mots		nb tours		nb mots	
S T	3 863 7 282	11 145	46 859 72 903	119 762	4 062 371	4 433	66 725 12 356	79 081
P + S P + T	3 819 7 698	11 517	30 293 108 278	138 571	4 098 382	4 480	33 686 4 156	37 842
Total	22 662		258 333		8 913		116 923	

TABLE 5.3 – Décomposition du corpus en sous-corpus, en nombre de tours de parole et nombre de mots, en fonction du type d'interlocuteurs : S (schizophrène), T (témoin), P + S (psychologue avec un schizophrène), P + T (psychologue avec un témoin)

Chapitre 6

Modélisation formelle des entretiens

Sommaire

On s'interroge régulièrement sur l'adaptibilité à des problématiques concrètes des travaux de linguistique informatique formelle. Au delà de l'intérêt de la modélisation pour elle-même se pose la question de sa validation d'un point de vue cognitif. Est-ce que les modèles proposés rendent effectivement compte de problèmes réels ou sont-ils la mise en œuvre de propriétés formelles justifiées par des analogies avec le monde réel ? Il ne s'agit pas ici d'apporter de réponse définitive. Cependant, il est pour le moins intéressant, sinon nécessaire, d'évaluer les modèles, sans toutefois leur donner un rôle de démiurge.

La question sous-jacente des travaux initiaux sur la représentation du discours schizophrénique était double : d'une part identifier des dysfonctionnements psycho-linguistiques chez des patients atteints de schizophrénie, et d'autre part travailler à définir l'« état de folie » communément accepté pour cette pathologie. Du point de vue épistémologique la folie est un concept qui oblige à définir le monde réel, le processus communicationnel en œuvre dans le langage et le positionnement de chaque réalité dans ce processus.

6.1 Discontinuités décisives et relations de discours

Nos travaux ont consisté dans un premier temps à former un corpus d'entretiens avec des schizophrènes ainsi qu'avec un groupe témoin appareillé et à identifier - comme l'ont montré (Musiol et Trognon 1996) - des discontinuités décisives dans ces discours. Comme nous l'avons introduit, elles sont définies comme présentant une rupture logique sur plusieurs tours de parole pendant l'échange. Seul le sous-groupe des schizophrènes paranoïdes [1] faisait apparaître ce type de discontinuités. Ainsi, la notion de discontinuités dialogiques est un indice psycho-linguistique discriminant dans l'identification d'une pathologie. Cette idée s'inscrit dans la continuité des travaux de (Chaika 1974) et (Fromkin 1975) qui se sont intéressés à l'impact de la pathologie sur la capacité langagière de ces populations.

Afin de passer à l'étape de la modélisation formelle de ces analyses, nous avons travaillé à représenter les entretiens à l'aide de la *Segmented Discourse Representation Theory* (SDRT) (Asher et Lascarides 2003). Cette théorie se base sur la DRT (Kamp

1. Terminologie définie selon le (DSMIV 1994)

et Reyle 1993). L'un des arguments en faveur de la DRT était qu'elle proposait un algorithme de résolution des anaphores pronominales [2] de manière intrinsèque. Cet algorithme apparaît malheureusement trop flexible dans certains cas, ce qui a conduit à l'introduction de la SDRT. L'exemple 6.1 est une simplification du célèbre exemple utilisé pour illustrer la SDRT (Asher et Lascarides 2003). Le discours construit ici une description d'une situation, où il n'est pas possible de poursuivre avec l'exemple 6.2. Le *It* devrait faire référence au « saumon », ce qu'un locuteur natif de l'anglais ne peut pas comprendre. La DRT permet à tord, elle, d'accéder à la variable qui représente l'entité saumon.

(6.1) *John experienced a lovely evening last night.*
John a eu un agréable diner hier soir.
He had a fantastic meal.
Il a eu un diner fantastic.
He ate salmon.
Il a mangé du saumon.
He devoured lots of cheese.
Il a dévoré une montagne de fromage.

(6.2) *It was a pink one.*
C'en était un rose.

Pour palier ce problème, la SDRT augmente la DRT d'une méta-représentation des relations de discours. L'exemple 6.1 est alors représenté par la structure de la figure 6.1. La représentation se bâtit à partir de relations de subordination (relations verticales) et de coordination (relations horizontales). La SDRT utilise cette structure pour définir les lieux possibles de rattachement de la suite d'un discours. Elle suppose qu'il s'agit de la partie qui est le long de la frontière droite. Dans notre exemple, l'intérêt réside dans la relation de narration entre « *He ate salmon* » et « *He devoured lot of cheese* ». Cette relation implique que le saumon n'est plus le long de la frontière droite, et donc que le saumon n'est plus accessible pour une référence ultérieure. Dans (Amblard et Pogodalla 2014), nous positionnons la DRT par rapport aux grammaires de Montague, et la SDRT par rapport à la DRT.

Au début de notre travail sur le discours schyzophrénique, nous avons dû répondre à plusieurs interrogations. Tout d'abord comment appréhender le type d'échanges en œuvre dans les entretiens. Il ne s'agit pas à proprement parler d'un discours puisque mettant en jeu deux intervenants, mais pas non plus d'un dialogue, puisque la fonction discursive du psychologue n'est pas naturelle. Il intervient pour maintenir l'échange en faisant parler le patient. Il aide le schizophrène à verbaliser et construire une pensée complexe. Il nous apparaît que les modèles sémantico-pragmatiques comme la SDRT sont alors tout à fait pertinents pour la modélisation. On se demande aussi, dans un mouvement inverse, comment la modélisation des phénomènes pathologiques permet d'interroger la validité cognitive de ces modèles.

2. Ce problème est celui d'identifier à qui un pronom fait référence parmi les entités précédemment introduites dans un discours.

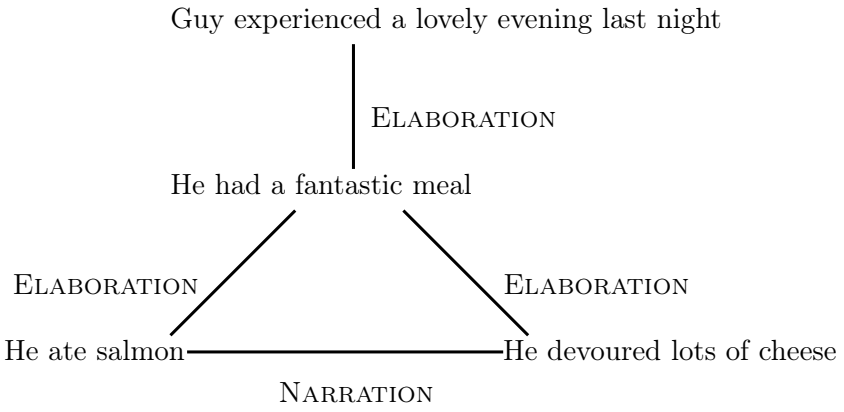


Figure 6.1 – Représentation en SDRT de l'exemple 6.1 (Asher et Lascarides 2003)

Plusieurs postulats se sont révélés nécessaires : d'abord que le patient schizophrène et le psychologue n'ont pas la même représentation mentale de l'interaction, au contraire de deux intervenants dans un dialogue qui cherchent à construire une représentation commune ; ensuite que l'irrationalité apparente chez le schizophrène pouvait aisément être placée à différents niveaux. Nous avons choisi de suivre le principe de charité de (Quine 1960) qui nous conduit à admettre que le patient est cohérent pour lui-même au moment où il intervient dans l'interaction. Ainsi lorsqu'un patient schizophrène énonce une phrase, il la considère comme valide. Nous ne remettons pas en question cette vérité. Ceci implique que les prises de parole du patient sont alors logiques (au sens tout à fait classique) au moment où elles sont réalisées. Les incohérences qui apparaissent ne dépendent donc pas du seul niveau sémantique mais sont transférées à un niveau de description plus abstrait de représentation comme la planification du discours.

De fait le patient a une production acceptable au niveau sémantique et il réalise des erreurs au niveau pragmatique. Mais ce n'est pas la vision de l'interlocuteur ordinaire qui lui cherche à interpréter normalement la production au niveau pragmatique et constate l'apparition d'erreur qu'il situe alors au niveau sémantique. Des contradictions apparaissent et l'interprétation n'est plus possible en l'état. Cette présentation implique de construire deux représentations différentes de l'échange qui dépend du point de vue adopté. La figure 6.2 rassemble ces éléments.

6.2 Représentation formelle de la discontinuité décisive

Nous avons obtenu une première série de résultats à différents niveaux, (Amblard, Musiol et Rebuschi 2011 ; Amblard, Musiol et Rebuschi 2012). Nous avons identifié comment la SDRT permettait de capter les interactions dans la modélisation des relations pour la rhétorique. Il faut cependant nuancer ce propos car une unité thématique entre les

Interprétation du discours par :	
Sujet normal (à la 3e personne)	**Schizophrène** (à la 1e personne)
hypothèse : correction pragmatique ⇓ déviance sémantique	déviance pragmatique ⇑ **hypothèse :** correction sémantique
contenu contradictoire : *apparence* de contradiction	contenu cohérent : *possibilité d'interpréter*

FIGURE 6.2 – Interprétation du discours en fonction du locuteur

différents tours de paroles doit être vérifiée au cours de l'échange. Elle peut aisément se calculer grâce aux formules logiques présentes dans la représentation en DRT. Nous avons donc étendu la SDRT pour conserver les arbres rhétoriques et nous l'avons augmentée de boîtes thématiques représentant la portée des thèmes en jeu dans l'interaction. Ensuite, nous avons analysé chaque extrait discontinu pour comprendre la partie dysfonctionnelle.

La figure 6.3 reprend un extrait porteur d'une discontinuité d'un entretien entre un patient schizophrène et un psychologue. Dans cette transcription, les tours de parole du patient sont identifiés par un B et ceux du psychologue par un A. Chaque tour est numéroté dans l'ordre d'apparition. Cette version originelle de la transcription portait des marques phonologiques simplifiées (avec des flèches), qui ont été normalisées par la suite. Nous avons mis en avant en gras les éléments utiles à l'interprétation de la discontinuité, uniquement pour cette présentation. On peut constater qu'il est difficile d'interpréter simplement la fin de l'échange. Nous reviendrons sur les éléments caractéristiques de cet échange.

Afin de rendre compte de ces phénomènes, il convient d'interroger les logiques en œuvre, comme nous l'avons fait dans (Rebuschi, Amblard et Musiol 2014) où nous discutons de leur adéquation avec les interprétations. Il faut noter que ce type de travaux s'inscrit dans une voie significativement différente de celles actuellement portées sur l'interprétation de la pathologie pour qui l'explication du dysfonctionnement est supposée provenir de l'expression d'un gène ou de la structure du réseau de neurones en action chez le patient, (Bloom 1993 ; Petronis 2004). *A contrario* nous admettons ici que d'autres aspects du langage et de la psychologie sont impliqués pour comprendre/expliquer la pathologie et nous adoptons une interprétation du discours en troisième personne.

Nous obtenons pour un même échange deux points de vue spontanés sur la conversation : celui du psychologue qui fait apparaître un contenu contradictoire, et celui du schizophrène qui ne respecte pas les constructions discursives mais présente un contenu sémantique cohérent (pour lui). Cette vision de l'échange induit deux éléments importants, d'une part qu'il faut construire deux représentations différentes en parallèle, et d'autre part que le problème d'interprétation n'apparaît pas au niveau sémantique, mais

B124 Oh ouais (↑) et pis compliqué (↓) et c'est vraiment très très compliqué (→) **la politique** c'est quelque chose quand on s'en occupe **faut être gagnant** parce qu'autrement quand on est perdant c'est fini quoi (↓)
A_{125} Oui
B_{126} J. C. D. **est mort**, L. **est mort**, P. **est mort** euh (...)
A_{127} **Ils sont morts parce qu'ils ont perdu à votre avis** (↑)
B_{128} Non ils gagnaient mais si ils sont morts, **c'est la maladie** quoi c'est c'est (→)
A_{129} Ouais c'est parce qu'ils étaient malades, c'est pas parce qu'ils faisaient de la politique (↑)
B_{130} Si enfin (→)
A_{131} Si vous pensez que c'est parce qu'ils faisaient de la politique (↑)
B_{132} Oui tiens oui il y a aussi C. **qui a accompli un meurtre là** (→) il était présent lui aussi qui est à B. mais enfin (→) c'est encore à cause de la politique ça

FIGURE 6.3 – Extrait de la transcription d'un entretien du corpus avec discontinuité

bien à un niveau pragmatique. Nous avons alors formulé deux hypothèses :

1. Les schizophrènes sont logiquement consistants. Cela implique que les ruptures ne peuvent intervenir qu'au niveau pragmatique, sur le processus de construction de la représentation de la conversation (sur les relations rhétoriques des SDRS), comme nous l'avons présenté dans la figure 6.2.
2. La sous-spécification (ambiguïté) est centrale dans la rupture. Un choix linguistique (la résolution d'une ambiguïté) n'est jamais définitif pour les schizophrènes.

Nous avons travaillé sur les corpus de La Rochelle et de Lyon, afin d'identifier manuellement l'ensemble des discontinuités présentes. Ce travail a été publié dans (Amblard, Musiol et Rebuschi 2014) qui les reprend exhaustivement. Nous avons montré que les discontinuités décisives correspondaient à des usages non-usuels en SDRT. Deux phénomènes se présentent : la remontée à travers la structure de représentations, sans consistance de la partie précédemment explorée et la rupture de la frontière droite. Cette dernière propriété est fondamentale dans la définition de la SDRT et permet de limiter les lieux de rattachement de la suite du discours. On remarque que, si la frontière droite permet d'améliorer le calcul, elle n'a pas par définition de réalité cognitive. Le fait qu'elle puisse être en œuvre dans ce phénomène pathologique permet d'en valider cognitivement l'existence.

Nous avons représenté chacun de ces extraits en suivant la SDRT et en l'augmentant d'une information thématique. Ce dernier point apparaît sous forme de boîte dans la figure 6.4. Par ailleurs, nous utilisons les relations issues de la SDRT, présentées dans la table 6.1.

En reprenant l'extrait présenté dans la figure 6.3, on voit apparaître dans la représentation deux thèmes : le premier porte sur la mort symbolique (la mort en politique)

Relations horizontales	Relations verticales	Relations obliques
Narration	Élaboration	Question
Réponse	Élaboration : explanation	Question : conduite
Réponse phatique	Élaboration : prescription	Question méta
Continuation	Évaluation	Élaboration requirement
et illustration	Phatique	conduite
	Coutre-élaboration	
	Justification	

TABLE 6.1 – Relations pragmatiques utilisées dans nos représentations

et le second sur la mort littérale. Les deux thèmes sont directement liés bien qu'ils expriment des réalités très différentes. La figure 6.4 présente la représentation en SDRT de cet extrait.

Les tours de parole B_{124} et A_{125} introduisent la première partie sur le thème de la mort symbolique. B_{126} amorce un nouveau thème sur la mort littérale, ce qui conduit à demander confirmation du changement de thème et obtenir sa confirmation. Si l'on n'accepte pas le changement de thème dès B_{126}, il devient évident avec B_{128}. La relance sous la forme d'une question du psychologue en A_{129} permet de conclure sur cette ambiguïté. De manière surprenante, le schizophrène au lieu d'abonder dans le sens de l'interaction revient sur cette interprétation et lui donne le premier sens utilisé. En cela il revient en arrière pour rattacher B_{130} (sinon B_{130} au moins de manière certaine B_{132}) à l'intérieur de la structure en A_{125}. Or, cette opération n'est pas possible en SDRT puisqu'il s'agit d'une rupture de la frontière droite.

L'analyse pourrait s'arrêter là, mais une fois que la conversation s'est stabilisée sur la première interprétation de la mort, le schizophrène change dans le même tour de parole en repassant à une interprétation de la mort littérale pour revenir à la mort symbolique. On est alors en présence d'une double tentative de rupture de la frontière droite, ce qui rend la reconstruction pragmatique difficile et donc l'interprétation non-acceptable.

On trouve par ailleurs d'autres exemples dans la modélisation, porteurs d'un autre phénomène : une inconsistance. S'il paraît possible de changer de thématique soudainement, un problème de cohérence apparaît lorsque l'interaction précédente reste en attente d'une information supplémentaire pour clore la structure en cours de développement. Une solution est toujours possible si on accepte que le rattachement sémantique puisse se faire plus haut dans la structure, et, dans le pire cas, tout en haut de la structure. Si la sous-structure reste ouverte, l'interprétation générale n'est alors pas acceptable.

L'extrait repris dans la figure 5.2 est modélisé par la représentation de la figure 6.5. On y trouve trois utilisations du terme « perdre », qui sont mises en gras dans la figure 5.2. La relance du psychologue en V_{87} [3] s'appuie sur ce terme pour éviter de s'enfermer dans une

3. Dans cette transcription, le psychologue est identifié par la lettre V.

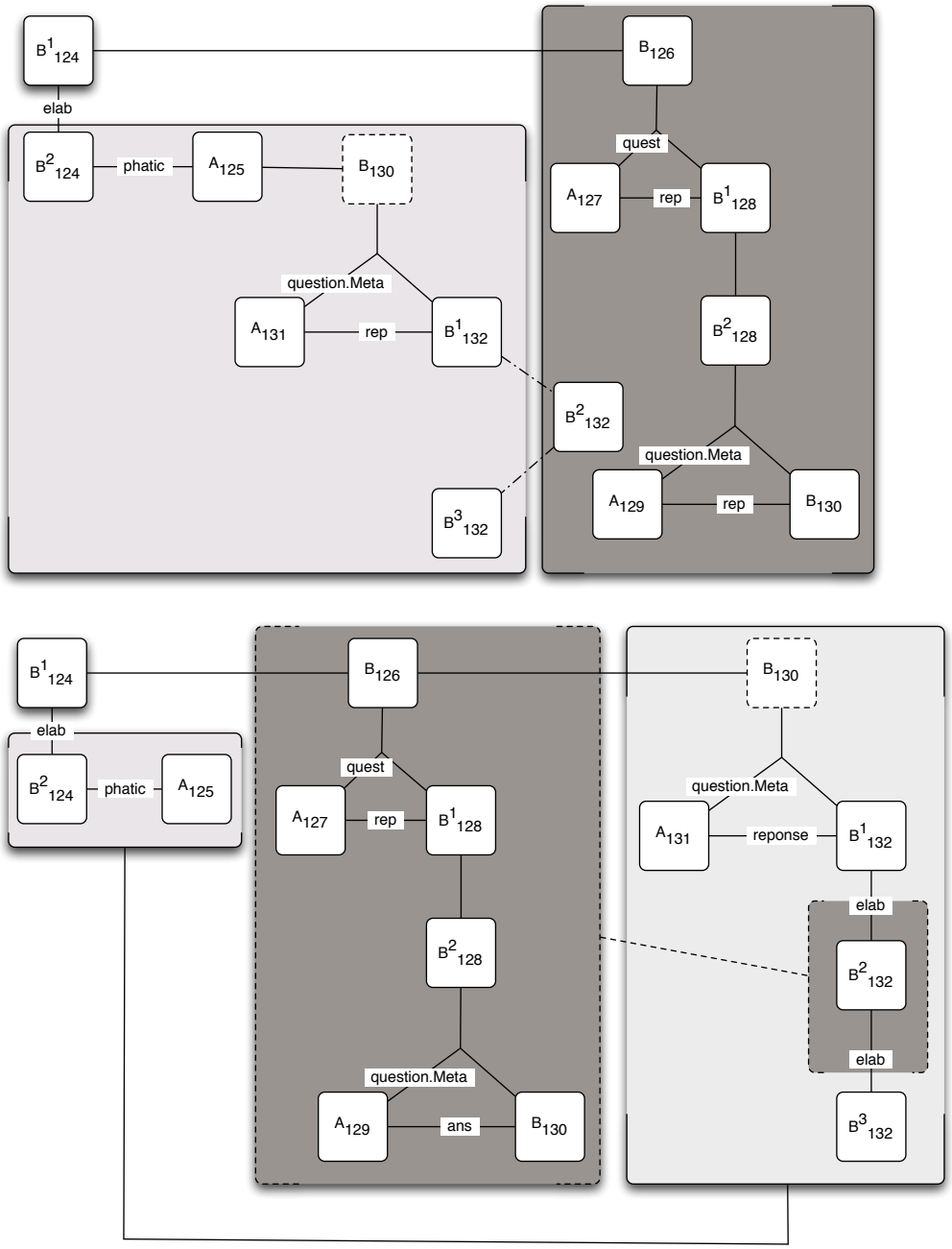


FIGURE 6.4 – Représentations inspirées de la SDRT de l'échange de la figure 6.3. La première est celle du schizophrène et présente des ruptures de la frontière droite, la seconde correspond à celle du psychologue, qui cherche à réparer l'échange pour le rendre interprétable.

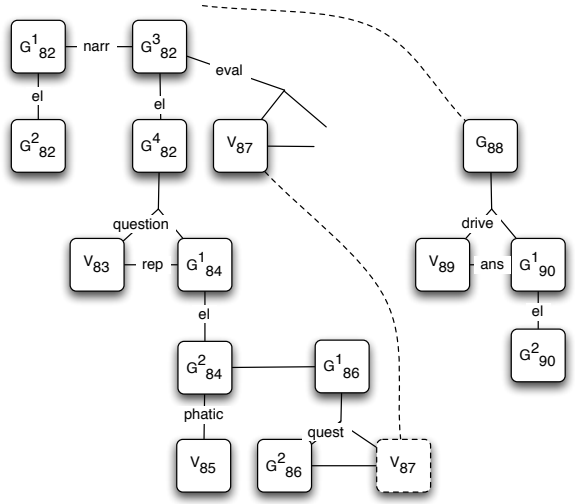


FIGURE 6.5 – Représentation inspirée de la SDRT de l'échange de la figure 5.2

interprétation. Cette montée semble amener le schizophrène à réaliser une remontée dans la structure mais sans s'arrêter au bon niveau de rattachement et donc sans pouvoir le positionner correctement. Ce phénomène implique une réinterprétation de « perdre » mais avec un second sens.

Nous supposons que les deux interlocuteurs tentent de répondre à leurs attentes mutuelles. Dans cette analyse, le déplacement de la rationalité chez le patient schizophrène provient de la relation linguistique qu'il entretient avec son contexte, et donc de la relation qu'il entretient avec son interlocuteur. Les troubles de la pensée qui apparaissent dans le langage interrogent la modélisation sémantique et pragmatique. Par extension nous travaillons également à l'identification de critères spécifiques pouvant être transférés dans la relation entre discours et cognition et sur la question de l'existence de signes pathognomoniques [4] de la maladie mentale. Il ne faut pas perdre de vue que ces approches entretiennent une vision normative du comportement. En effet, nous supposons ici un usage standard de la langue à partir duquel on identifie une déviation supposée manifeste d'un symptôme, mais il nous faudrait revenir sur la définition de la normalité de la performance langagière.

4. se dit d'un symptôme caractéristique d'une seule maladie.

Chapitre 7

Annotations du corpus

Sommaire

Notre analyse est du type sémantico-pragmatique comme nous l'avons abordé dans le chapitre précédent. Il nous a semblé nécessaire de vérifier si d'autres niveaux linguistiques étaient également porteurs d'indices d'une défaillance dans la production langagière (et par extension d'une défaillance des processus cognitifs les portant). On considère que la transcription est la première couche d'annotation de la ressource. Elle conserve cependant une fonction particulière puisque les autres annotations sont réalisées à partir de cette première. Nous nous sommes alors intéressés à l'obtention de ces informations en mobilisant des outils du traitement automatique des langues d'une part, et en mettant en place des campagnes d'annotations manuelles (pour les niveaux ne possédant pas d'outil produisant une qualité suffisante). Dans ce chapitre nous revenons sur notre étude de l'apparition des phénomènes de disfluence et de la répartition des catégories morpho-syntaxiques (*POS*) à partir d'outils automatiques. Ces aspects ont donné lieu à l'implémentation d'un outil, `SLAMtk`, sur lequel nous reviendrons dans la section 7.2. Pour les autres niveaux dépourvus d'outils automatiques nous avons travaillé à la définition de campagnes d'annotations manuelles pour identifier les discontinuités décisives, et définir leur modélisation formelle. Nous reviendrons sur les motivations et les réalisations dans la section 7.3. Avant d'arriver à la production de grands volumes d'annotations par les outils ou la production manuelle de peu d'annotations de bonne qualité, nous avons fait face au problème de la diffusion de la ressource et l'impossibilité de l'anonymisation.

7.1 L'impossibilité de l'anonymisation

Motivés par la volonté de produire des annotations manuelles pour trouver un concensus large, nous avons dû évaluer la qualité de l'anonymisation que nous pouvions assurer sur la ressource. Nous avons choisi de produire une version anonymisée à partir de MEDINA (Grouin et Zweigenbaum 2013), un outil dédié à ce type de tâche. Suite à des complexités administratives pour obtenir les droits d'utilisation, nous n'avons pas pu en disposer. La solution utilisée est moins efficace, mais elle a permis de répondre à une partie de nos besoins.

spk1 Et donc euh j' avais j' ai pendant trois trois quatre ans **j' avais commencé des études** j' ai fait un peu différentes choses parce que

...

spk1 Euh **dans une école d' ingénieur à Ville1 dans dans le nord** euhh et donc euhh euhh ouais donc j' ai je c' est là où j' ai commencé à être malade en fait juste [...]

spk1 donc du coup ben là c' est je j' ai j' ai repéré deux trois le le c' était quand même assez stressant euh **la la prépa**

spk2 Mmh mmh

spk1 donc euh donc du coup ouais euh et bon pour euh en ce qui concerne les études donc du coup après j' ai j' ai arrêté le le le le l' école d' ingénieur enfin la prépa **je suis revenue à Ville2**

spk2 Mmh mmh

spk1 **j' ai fait euh une une une fac de de maths** je suis allé en fac de maths

FIGURE 7.1 – Entretien anonymisé mais pour lequel de nombreux éléments de réidentification persistent

Nous nous sommes alors penchés sur le développement de scripts Python qui facilitent le travail de désidentification. Il s'agit de trouver des entités nommées par des expressions régulières qui respectent les règles graphiques. Une catégorie est assignée à chaque entité et le script substitue une marque sémantiquement vide à toutes les instances présentes dans l'entretien. À partir des actes de langage, on conserve sa cohérence interne. Ces scripts sont interactifs et ils permettent à l'utilisateur d'ajouter des éléments à désidentifier.

Nous nous sommes résolus à penser, comme d'autres avant nous (Eshkol-Taravella et al. 2014), que s'il nous était possible de cacher le prénom et le nom des personnes, les locuteurs exprimant de nombreuses informations tant sur leur histoire personnelle que sur leur famille, voire sur leur localisation, il nous était impossible de garantir un véritable anonymat.

La figure 7.1 présente la transcription d'un entretien avec ces problèmes. Bien que les noms de ville ou de lieu aient été remplacés, et que l'échange soit relativement court, nous pouvons constater qu'il permet d'obtenir de nombreuses informations biographiques. On sait qu'il s'agit d'une personne qui a suivi des études d'ingénieur dans le Nord, sans terminer son cursus. Le nombre d'écoles est limité, plus encore avec l'indication géographique. Par ailleurs, elle s'est inscrite dans une faculté de mathématiques. Rien que ces éléments permettent de considérablement augmenter les possibilités d'identification. La suite de l'entretien, bien évidemment pas présentée dans ce document, fournit encore d'autres éléments significatifs. Il ne s'agit bien entendu pas ici de dévoiler l'identité du patient, mais bien de donner l'intuition au lecteur qu'il n'est pas possible de garantir l'anonymat des personnes, tout en conservant la cohérence de leur dialogue. C'est pourtant cette

cohérence qui est à la base de leur discours et donc l'objet de notre étude. On peut argumenter qu'il est possible d'obtenir l'autorisation éclairée des patients. Cependant, la nature des entretiens les pousse à délivrer des éléments non seulement sur eux, mais aussi sur leur famille et leurs proches. Les informations ont souvent un caractère personnel, et nous ne possédons pas d'autorisation éclairée pour les diffuser (Amblard, Fort, Musiol et al. 2014).

Pour palier ce problème, nous avons produit une version de la ressource où tous les tours de parole sont mélangés aléatoirement. On perd la narration de chaque entretien, ce qui rend impossible de reconstruire une histoire, une temporalité ou une géographie significative. Il est ainsi possible d'étudier les productions quantitatives comme l'apparition de phénomènes de disfluence, mais il ne fait plus sens de travailler sur l'interprétation discursive. Cette solution peut être mise en œuvre sur une partie du projet.

Ce constat interroge la pérennité de ce type de recherche. Il est en effet évident qu'il y aurait nécessité à rendre public le contenu d'une telle ressource, non pas par voyeurisme, mais bien parce qu'il est difficile d'avoir accès à des corpus d'usage pathologique de la langue. Des études multiples et fines de tels corpus permettraient d'avancer plus efficacement sur ce type de questions relatives à la pathologie. Il semble également normal de rendre publiques des données qui ont été financées par de l'argent public, ne serait-ce que pour la duplication des résultats de la recherche.

Il faut aussi considérer la volonté institutionnelle qui oblige à respecter la confiance donnée par les patients dans la recherche qui leur garantit une anonymisation des données.

7.2 SLAMtk : Distager et MElt

Nous avons entrepris d'annoter automatiquement la ressource grâce à des outils du TAL. Les deux types d'annotations sur lesquels nous nous sommes concentrés sont l'identification des phénomènes de disfluences et des catégories morpho-syntaxiques. Par extension, nous avons travaillé sur une analyse textométrique. Ce dernier aspect a été plus particulièrement porté par Karën Fort. Dans cette section nous revenons d'abord sur les outils utilisés, puis sur les résultats obtenus. Tous ces éléments ont fait l'objet de plusieurs publications, notamment (Amblard et Fort 2014 ; Amblard, Fort, Demily et al. 2015).

7.2.1 Annotations des disfluences : Distagger

Le corpus étant constitué de transcriptions de l'oral, il est normal d'y trouver des marques de disfluences. Les phénomènes de disfluences sont, de manière traditionnelle, des réalisations de l'oral qui rompent la continuité syntaxique. Nous avons utilisé l'outil d'identification automatique `Distagger` (Constant et Dister 2010). L'outil est libre et présente de bonnes performances sur un corpus de référence de 22 476 mots et 1 280 disfluences, à 95,5 % de F-score (précision de 95,3 %, rappel 95,8 %).

`Distagger` permet d'identifier des réalisations de natures différentes, pour lesquelles quatre restent prédominantes dans les corpus oraux : les interjections d'hésitation (« euh »),

```
{S}{#25,.IGN+slot} {spk2,.IGN+speaker}
Bof {euh,.IGN+EUH} {/,.IGN+short_pause} quand j'ai envie {de /,.IGN+REP}
de faire un tour je fais un tour si je vois que j'ai pas envie
{quand il c/,.IGN+REPFRAG} quand il pleut quand il fait mauvais eh ben je
reste à la maison
{S}
```

FIGURE 7.2 – Annotations d'un tour de parole de la ressource par `Distagger` (annotations des interjections d'hésitation (« heu »), des répétitions, et autres informations spécifiques à l'outil)

les répétitions, les autocorrections immédiates et les amorces de morphèmes. Ces différentes réalisations sont définies comme suit (les exemples proviennent du corpus).

— Les interjections d'hésitation (« euh ») :

(7.1) moi ça m'est presque plus euh difficile et euh anti-naturel de parler

— Les répétitions sont entendues comme la reprise explicite et identique d'un mot ou d'un groupe de mots dans le contexte immédiat d'apparition. La répétition peut contenir ou être précédée d'un mot creux comme « oui », « non », ou un « euh » :

(7.2) j' arrive à être à être concentrée quand il faut faire quelque chose

— L'autocorrection immédiate est une variante de la répétition dans laquelle un trait morphologique peut varier (ce qui apparaît régulièrement avec les déterminants) :

(7.3) enfin je sais pas trop le les termes

— L'amorce est une interruption de morphème en cours d'énonciation. La fin du mot est marquée par un tiret :

(7.4) pis progressivement vous av- pouvez travailler sur votre concentration

Le résultat de l'annotation par `Distagger` sur les corpus produit sept étiquettes, dont deux particulières pour marquer les tours de parole et identifier les interlocuteurs. Par ailleurs, deux autres types apparaissent dans des volumes trop restreints pour être significatifs. Nous avons travaillé sur les trois étiquettes : $\{EUH\}$ pour les interjections d'hésitation, $\{REP\}$ pour les répétitions et $\{CORR\}$ pour les auto-corrections. Un exemple d'annotations par `Distagger` est présenté dans la figure 7.2.

Nous avons pu constater que les amorces sont mélangées avec les répétitions et les corrections. Comme nous nous sommes concentrés sur les volumes des disfluences et non sur leur ventilation en catégories, nous avons conservé cette version de la ressource. Les perspectives d'évolution nécessitent de revenir sur ce point pour affiner les résultats obtenus avec cet outil sans se contenter de travailler sur le *reparandum*[1].

1. Le reparandum est la partie de l'énoncé précédant la disfluence et qui doit être corrigée.

7.2.2 Annotations morpho-syntaxiques : MElt

Pour l'analyse morpho-syntaxique, nous avons choisi de travailler avec l'outil MElt (Denis et Sagot 2009). Il s'agit d'un analyseur (*tagger*) lui aussi librement disponible. Il est basé sur des méthodes d'apprentissage, en particulier des perceptrons multicouches. Une version de l'outil a été entraînée sur le corpus TCOF-POS (Benzitoun, Fort et Benoît Sagot 2012) du français parlé et utilise le lexique Le*fff* (Benoit Sagot 2010). Les modèles nous sont apparus suffisamment proches de nos données pour être légitimement mobilisés. L'outil obtient de bonnes performances, 97,61 % d'exactitude, ce qui le met au niveau des autres outils existants.

L'outil est appliqué sur chacun des entretiens, ce qui alourdit le temps de traitement car le temps de chargement des modèles est assez important. Les fichiers sont divisés en deux parties pour que l'outil soit appelé sur le contenu textuel, sans autre marque spécifique du corpus, la seconde partie contenant les informations permettant de rattacher le texte à sa position et ses informations dans la ressource. La ressource augmentée des annotations en catégories morpho-syntaxiques est reconstruite *a posteriori*.

L'exemple 7.5 présente le résultat de l'annotation de MElt sur un extrait du corpus. L'outil commence par découper le texte en mots, et associe à chaque mot une catégorie morpho-syntaxique et son lemme. MElt utilise le caractère $*$ pour les catégories morpho-syntaxiques non-identifiées ou les lemmes non-présents dans le lexique.

(7.5) Voilà alors peut-être vous pouvez m'e/ m'expliquer
Voilà/FNO/voilà alors/ADV/alors peut-être/ADV/peut-être vous/PRO :cls/vous pouvez/VER :pres/pouvoir m'/PRO :clo/me e/ADV/*e //MLT/*/ m'/PRO :clo/me expliquer/VER :infi/expliquer

Les étiquettes des catégories morpho-syntaxiques produites se déduisent simplement : PRO signifie pronom, ADV, adverbe, VER verbe, *etc.*

7.2.3 Analyse textométrique : TXM

L'analyse textométrique est réalisée avec l'outil TXM (Heiden, Magué et Pincemin 2010), outil lui aussi librement disponible qui inclut des fonctionnalités d'analyse statistique (*via* le logiciel R). Outre sa facilité d'utilisation, TXM présente un avantage décisif par rapport à des logiciels d'analyse statistique générique : il offre un accès direct au contexte, ce qui permet d'affiner les résultats quantitatifs par une analyse qualitative manuelle. Nous examinons les contextes des lemmes présentant des spécificités élevées, afin de vérifier à quoi ceux-ci correspondent dans le corpus.

TXM utilise un étiquetage des corpus en catégories morpho-syntaxiques obtenu avec TreeTagger (Schmid 1994). Les résultats de MElt étant de meilleure qualité (environ deux fois moins d'erreurs), nous importons notre annotation précédente. TXM nous permet d'évaluer l'apparition d'un terme dans des sous-parties du corpus. Ces comparaisons créent une vision globale de la production des schizophrènes par rapport aux témoins. Pour valider ces résultats, nous calculons les spécificités (Lafon 1980) de chaque lemme, ce qui permet de prendre en compte les déséquilibres entre sous-corpus.

Nous calculons par ailleurs un indice de diversité lexicale qui correspond au ratio du nombre de lemmes par rapport au nombre total de formes différentes. Nous calculons également la richesse lexicale (RL) dans les sous-corpus des schizophrènes, témoins et psychologues. La RL est le ratio du nombre de lemmes par rapport au nombre total de formes. La RL se différencie du *type token ratio* (TTR) par le fait que nous utilisons les lemmes et non les types (qui sont des formes). La diversité lexicale permet de minimiser l'impact des mots très utilisés dans le calcul de la richesse lexicale. Plus la valeur est proche de 1, plus l'interlocuteur utilise de termes différents, indépendamment de leurs dérivations morphologiques.

7.2.4 Mise en œuvre des outils : `SLAMtk`

`Ditagger` et `MElt` sont appelés en ligne de commande, directement sur les documents textuels auxquels nous apportons des métadonnées. Nous avons donc implémenté une série de scripts en `Python` pour prétraiter le corpus et lui appliquer `Ditagger` et `MElt`. Par ailleurs, `TXM` est utilisé manuellement.

L'ensemble des scripts permettant de traîter ces différentes annotations est rassemblé dans un outil appelé `SLAMtk`. Cet outil a fait l'objet d'une première version, puis le code a été repris pour rassembler plusieurs éléments. L'implémentation est en cours de simplification afin de le distribuer librement. La quantité de traitements proposés étant importante, le développement en est assez complexe.

Par ailleurs, nous avons automatisé les analyses et produit différentes représentations à partir des données d'entrée. La section suivante reprend à la fois les illustrations et les résultats des calculs.

7.2.5 Résultats des analyses

L'ensemble des analyses a fait l'objet d'une publication (Amblard, Fort, Musiol et al. 2014) qui expose en détail les différents éléments. Nous n'en reprenons ici que les principaux aspects. Nous avons choisi de considérer que des résultats faisaient sens en fonction d'un calcul de significativité. Le principe est détaillé dans (Boula de Mareüil et al. 2013).

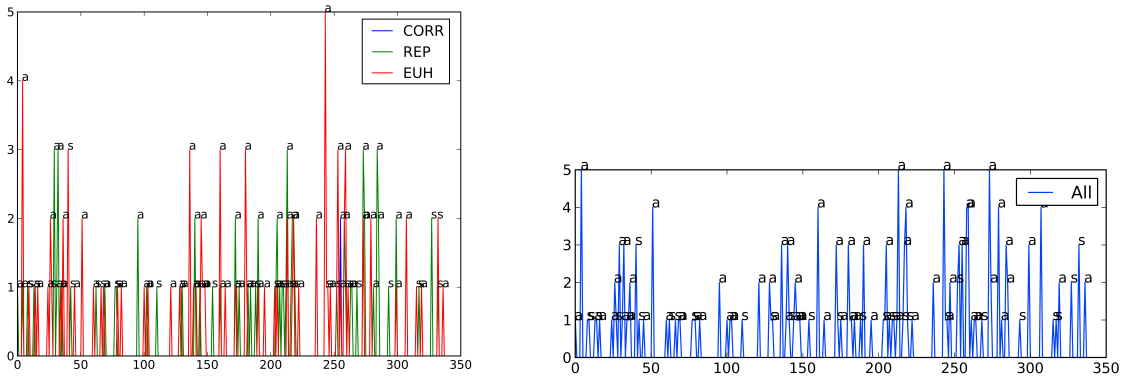
La significativité permet de calculer un indice de distribution en fonction du nombre de mots entre deux types d'interlocuteurs. Nous calculons systématiquement ces valeurs pour les appariements que nous pouvons construire avec les schizophrènes, les psychologues et les témoins. Nous cherchons grâce à la mesure à écarter des hypothèses de similarité de comportement en fonction d'une loi normale. Les résultats sont alors comparés à des valeurs spécifiques (en particulier 1,96 avec un risque d'erreur de 5%).

$$s = \frac{(p_1 - p_2)}{\sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}}$$

où :

— $p = (n_1 p_1 + n_2 p_2)/(n_1 + n_2)$;

— n_i est le nombre de mots [2] prononcés par la i^e catégorie d'interlocuteurs ;
— p_i est la proportion du phénomène attribuée à la i^e catégorie de locuteur.



(a) (b)

FIGURE 7.3 – Nombre d'étiquettes de disfluences par tour de parole pour un entretien. L'abscisse est la position du tour de parole dans l'entretien. Les tours de parole du psychologue sont notés par un s et ceux du schizophrène par un a.

L'analyse des disfluences a mis en évidence un schéma récurrent intéressant de comportement chez les schizophrènes. Ces derniers présentent un pic de stress en début d'entretien qui redescend avant de remonter au bout des deux-tiers de l'entretien, et ce, étonnamment, quelle que soit la durée de l'entretien. Sans avoir d'interprétation de ce phénomène, il est à corréler avec l'état de fatigue du patient. La figure 7.3 présente un exemple de la répartition des disfluences sur un entretien, et par type d'étiquettes en (a) et la somme des disfluences en (b).

	La Rochelle			Lyon		
	S	T	P	S	T	P
CORR	$0,0013$	$0,0007$	$0,0006$	$0,0004$	$9e-05$	$0,0001$
REP	$0,0211$	$0,0134$	$0,0174$	$0,0125$	$0,0078$	$0,0067$
EUH	$0,0369$	$0,0326$	$0,0282$	$0,0190$	$0,0089$	$0,0073$
Total	$0,0595$	$0,0468$	$0,0463$	$0,032$	$0,0168$	$0,0142$

TABLE 7.1 – Répartition des étiquettes de `Distagger` dans les sous-corpus, normalisée par rapport au nombre de mots (T = témoins, S = schizophrène, P = psychologue)

Le résultat intéressant que nous avons obtenu est que les patients schizophrènes produisent plus de disfluences que les témoins ou les psychologues. La table 7.1 présente la répartition des étiquettes de `Distagger` dans différents sous-corpus. Bien que les valeurs

2. Chaque *token* compte pour un mot (y compris dans les phénomènes de disfluences).

	La Rochelle	Lyon
T - P	$0,42$	$3,23$
S - P	$10,68$	$19,42$
S - T	$10,28$	$16,04$

TABLE 7.2 – Significativité des disfluences entre les groupes d'interlocuteurs (T = témoins, S = schizophrènes, P = psychologues)

soient petites, une différence apparaît sur les totaux entre les schizophrènes et les autres intervenants. La variation entre les deux corpus provient du type de transcription réalisée dans chaque sous-corpus. Le calcul de la significativité, présenté dans la table 7.2 valide l'hypothèse que le comportement est particulier au sous-corpus des schizophrènes.

D'autre part, nous avons travaillé sur les résultats de `MElt` pour les catégories morpho-syntaxiques et les lemmes. La table 7.3 présente les différentes fréquences d'apparition pour les catégories morpho-syntaxiques. Aucune variation de valeur ne semble apparaître en première lecture. Les calculs de la richesse lexicale et de la diversité lexicale de la table 7.4 ne font pas non plus apparaître de comportement spécifique. Les schizophrènes ont une diversité lexicale et une richesse lexicale équivalentes aux autres catégories testées, ce qui est confirmé par le calcul de la significativité que l'on retrouve dans (Amblard, Fort, Demily et al. 2015).

		Ratio	Diff	VER	ADJ	ADV	NOM	DÉT	PRÉ	PRO	AUT
La Rochelle	T	10,98	40	524	83	711	297	234	218	617	785
	S	13,12	38	416	61	537	238	183	190	497	632
	P	13,02	39	575	91	795	318	247	247	634	738
Lyon	T	31,78	35	247	45	173	199	146	141	265	243
	S	17,32	37	378	58	243	266	201	182	440	498
	P	9,42	35	192	27	135	117	86	80	231	194

TABLE 7.3 – Ratio moyen du nombre de catégories par rapport au nombre de tours de parole par entretien, nombre moyen d'étiquettes différentes par entretien, et répartition moyenne des catégories morpho-syntaxiques en grandes catégories : VERbe, ADJectif, ADVerbe, NOM, DÉTerminant, PRÉposition, PROnoms et AUTres (T = témoins, S = schizophrènes, P = psychologues).

Nous renvoyons le lecteur intéressé à la lecture de (Amblard, Fort, Demily et al. 2015) qui détaille ces analyses, et l'analyse textométrique. Bien qu'ayant participé au travail autour de ces questions, elles ne sont pas reprises ici. Le point intéressant de cette étude est que, malgré les possibles biais qui apparaissent entre la prise ou non de médicaments, la diversité entre hommes et femmes ou la prise en compte du QI des

	La Rochelle		Lyon									
					H		F		Avec trait.		Sans trait.	
	RL	DL	RL	DL	RL	DL	RL	DL	RL	DL	RL	DL
T	0,04	0,68	0,11	0,73	0,15	0,76	0,14	0,74				
S	0,05	0,69	0,06	0,70	0,07	0,72	0,08	0,71	0,06	0,71	0,10	0,72
P	0,02	0,64	0,06	0,68								

TABLE 7.4 – Richesse lexicale (RL) et diversité lexicale (DL) pour les sous-corpus, avec données selon le sexe et selon la prise ou non de traitements pour les schizophrènes pour le sous-corpus Lyon (T = témoins, S = schizophrènes, P = psychologues)

patients, il apparaît que les schizophrènes ont une fluence langagière moins performante que les témoins ou les psychologues, mais une richesse lexicale syntaxique équivalente. Ils ont donc des difficultés dans la gestion du système articulo-phonatoire, mais pas dans les processus cognitifs de production langagière. Aussi, les discontinuités décisives que nous avons évoquées apparaissent bien à d'autres niveaux linguistiques, en particulier dans la sémantique et la pragmatique. Cet aspect valide nos hypothèses de départ. Plusieurs améliorations et extensions peuvent être apportées à l'outil `SLAMtk`, ces questions seront reprises dans le chapitre 9.

7.3 Campagnes d'annotations manuelles

Afin d'avoir une approche homogène des différents niveaux, nous pouvons considérer que les annotations précédentes, produites automatiquement, appartiennent à des campagnes d'annotations automatiques. Comme nous l'avons indiqué dans la figure 5.3, nous sommes également concernés par des annotations pour lesquelles il n'existe pas d'outil du traitement automatique des langues disponible.

Nous nous sommes intéressés à la définition et à l'animation de campagnes d'annotations sur la ressource. Il a été montré que le succès de ces campagnes dépend principalement de leur réutilisabilité qui est directement liée à la qualité de leur définition initiale (Fort 2012).

Une première étape consiste à identifier les extraits présentant des discontinuités décisives. En effet, leur identification est faite par des spécialistes de la psychologie qui maîtrisent la théorie sous-jacente à leur justification. Nous nous sommes demandés comment d'une part valider l'existence de ces discontinuités de manière empirique et d'autre part s'il était possible de définir plus simplement les éléments en jeu dans ces relations entre interlocuteurs.

Nous sommes donc partis du principe de faire réaliser une annotation de la ressource pour identifier les discontinuités, puis chercher s'il était faisable de mettre en avant un accord inter-annotateurs. Nous avons encadré trois groupes d'étudiants de master première

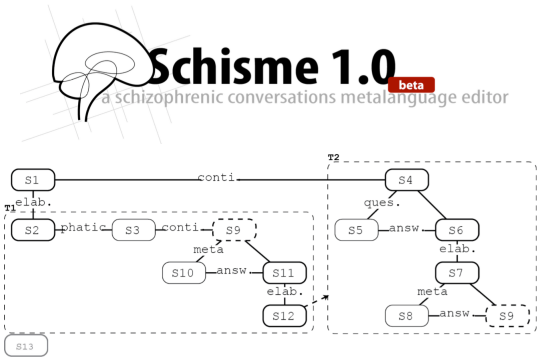


FIGURE 7.4 – Présentation graphique avec le logiciel Schisme

année qui ont travaillé sur ces questions. Le premier en 2011 a participé au développement d'une plateforme, `Shisme`, dont une copie d'écran est présentée dans la figure 7.4, qui se proposait d'être une interface pour l'annotation de ressource spécialisée dans notre contexte, en particulier à partir des schémas de la SDRT. Le résultat était très satisfaisant au niveau du rendu graphique, mais de nombreuses difficultés persistaient sur la gestion des annotations. Il s'est avéré plus pertinent de porter les résultats vers un logiciel dédié à l'annotation, `glozz` (Mathet et Widlöcher 2011), même si les couches graphiques mobilisables pour ce type de niveau étaient de bien moins bonne qualité. À partir de ces résultats, nous avons travaillé à la définition d'une campagne auprès d'une dizaine d'experts, principalement des étudiants de master de psychologie ayant suivi un enseignement sur les troubles du langages et de la pensée dans leur formation. Il est apparu que la phase de développement des outils avait été trop longue pour réussir à tester les interfaces avec un protocole dédié auprès d'annotateurs humains.

Un second groupe d'étudiants de master a travaillé en 2014 sur des problématiques liées à l'analyse en catégories morpho-syntaxiques, qui a donné des résultats moins pertinents que ceux précédemment présentés. Ce groupe a également proposé la définition d'une première campagne d'annotations. Dans ce cadre, ils ont rédigé un guide d'annotations pour l'identification des discontinuités décisives. Le guide est une présentation simplifiée de la tâche, qui doit être compréhensible par tous et ne doit pas expliciter les attendus de la recherche. Nous avons choisi 4 extraits du corpus, dont deux contenaient des discontinuités. Les étudiants ne connaissaient ni les fichiers contenant des discontinuités, ni leur position dans l'extrait. Ils ne pouvaient ainsi pas influencer les annotateurs. Les extraits contenaient de 45 à 60 tours de parole.

Les premiers annotateurs ont rapidement relevé que la tâche était beaucoup trop complexe pour réaliser l'annotation de tous les entretiens et ils se sont concentrés sur le premier. Par ailleurs, une difficulté d'organisation a imposé aux étudiants de solliciter des annotateurs non-spécialistes de la question. Cinq annotations d'un même extrait ont ainsi

pu être produites. Un annotateur a trouvé exactement la discontinuité, trois ont produit des représentations complexes dans la zone lui correspondant mais sans lui attribuer le statut de discontinuité et un annotateur n'a rien identifié. Il est encourageant de noter qu'un non-spécialiste a trouvé exactement la structure attendue, ce qui nous conduit à valider en grande partie le protocole expérimental. Quelques améliorations doivent être apportées pour accélérer la compréhension du guide d'annotations.

Finalement, il s'avère nécessaire de mieux anticiper l'analyse résultant des annotations en définissant plus finement la notion d'accord inter-annotateurs sur ces données. Il s'est également avéré nécessaire de développer des outils techniques pour récupérer les résultats et les traiter efficacement. Dans ce cas, les étudiants ont dû reconstruire à la main les représentations ce qui a un coût d'investissement non réaliste pour une campagne à grande échelle.

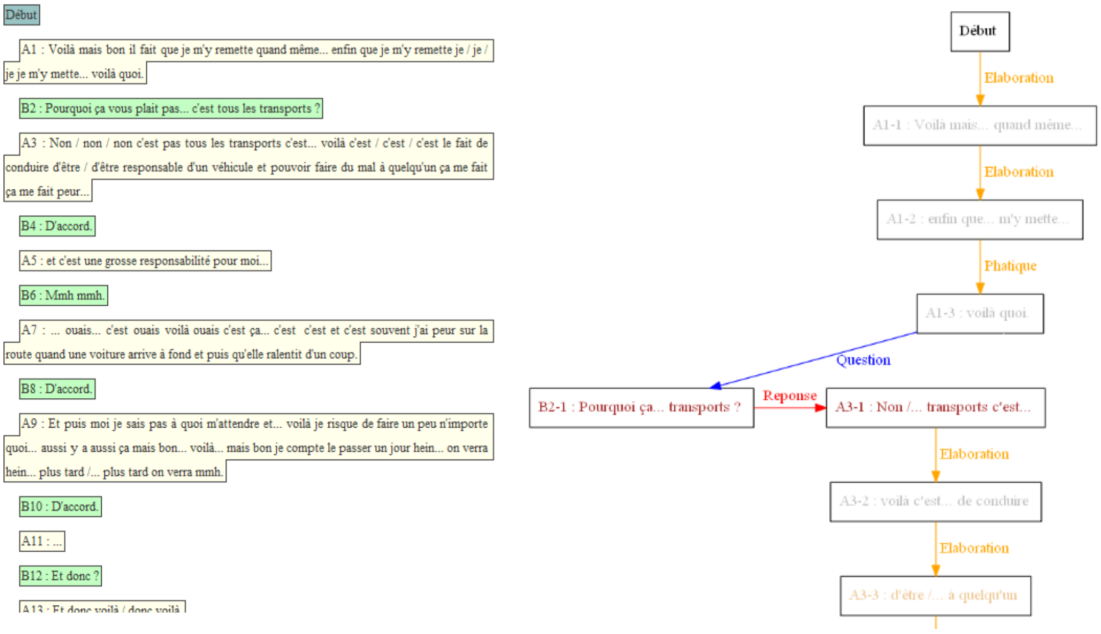
Un troisième groupe d'étudiants a travaillé en 2015 sur l'annotation, mais cette fois en SDRT. Ils ont donc repris la production du premier groupe sur l'annotation en SDRT, les résultats du second groupe sur la campagne d'annotations en discontinuité, le tout sur la plateforme `Glozz`. À cette fin, ils ont eux-aussi débuté par la rédaction d'un guide d'annotations. La première partie de la rédaction a conduit à faire des choix sur le type de représentation attendue, et de simplification des annotations possibles. Ils ont aussi demandé à ce qu'une définition du thème soit explicitement donnée par l'annotateur.

Par ailleurs, un prétraitement des fichiers s'est avéré nécessaire en construisant une version XLM augmentée de méta-informations, afin de procéder à des calculs statistiques automatiques *a posteriori*. Les étudiants ont fait le choix d'afficher les pré-annotations, comme la segmentation en tours de parole associant une même couleur à chaque intervenant. Ce résultat permet d'accélérer le travail des annotateurs et d'uniformiser les productions. Les annotateurs ne peuvent ajouter ni un caractère de plus, ni un de moins dans leurs segmentations. La figure 7.5 (a) présente le résultat de l'application de ces prétraitements, et donc un exemple de ce qui est présenté à l'annotateur (avant qu'il n'ajoute ses annotations structurelles).

L'annotateur a disposé d'un guide, d'une fiche de synthèse, d'une vidéo d'explication et d'une interface où les données sont prêtes à être analysées. Nous avons par ailleurs veillé à ce que les extraits ne soient pas trop longs pour ne pas décourager l'annotateur dans sa tâche. Nous avons demandé à ce que l'annotateur soit aidé avec une présentation sous forme d'arbre de sa production. La figure 7.5 (b) présente le résultat de l'affichage. L'annotateur disposait d'un bouton pour rafraichir l'affichage quand il le souhaitait.

Grâce à cet important travail de préparation, les étudiants ont pu faire passer leur test à 22 participants, principalement issus d'un cursus de psychologie, c'est-à-dire non-experts de la théorie linguistique, mais proches de la problématique générale du projet. L'organisation de ces campagnes d'annotations est lourde et ne peut être démultipliée.

Des données ont ensuite été automatiquement extraites des annotations, en particulier sur le nombre de thèmes utilisés, le nombre d'unités (segmentation interne de chaque tour de parole), ainsi que sur la répartition de l'utilisation des différentes relations. L'analyse montre que les annotateurs ont un comportement assez homogène sur l'ensemble des données. Il est par contre difficile de tirer des conclusions car l'échantillon d'annotateurs



(a) (b)

FIGURE 7.5 – (a) Version de la ressource à annoter en SDRT après prétraitements – (b) Résultat de l'annotation en SDRT par un annotateur

reste petit. Mais il faut remarquer que le temps moyen pour annoter les différents extraits choisis est d'environ 50 minutes.

La mise en place de ces différentes campagnes n'a pas donné lieu à des résultats spécifiques, mais a permis d'identifier clairement les besoins d'outils à déployer dans ce contexte. Une perspective est d'intégrer ces éléments dans l'outil `SLAMtk`. Nous considérons qu'il reste encore un travail conséquent de définition des méthodologies de post-traitements des données avant de lancer des campagnes à grande échelle.

Nous avons par ailleurs préparé des éléments de comparaison entre les différentes campagnes tests pour identifier les parties réutilisables de nos premières expérimentations. La figure 7.6 présente la complexité des campagnes d'annotations selon les critères introduits dans (Fort 2012). On constate que l'identification des discontinuités (en rouge) demande beaucoup plus d'investissement que celle pour la SDRT. Cet argument va dans le sens de notre constatation empirique où la campagne sur l'annotation en SDRT est bien plus réussie que celle pour les discontinuités.

Quoi qu'il en soit tous ces travaux ont grandement permis de faire avancer la réflexion sur l'annotation, et surtout aucun n'est venu s'inscrire en faux de nos hypothèses de départ quant à l'existence de discontinuités décisives et à la possibilité de les interpréter par des formalismes sémantico-pragmatiques.

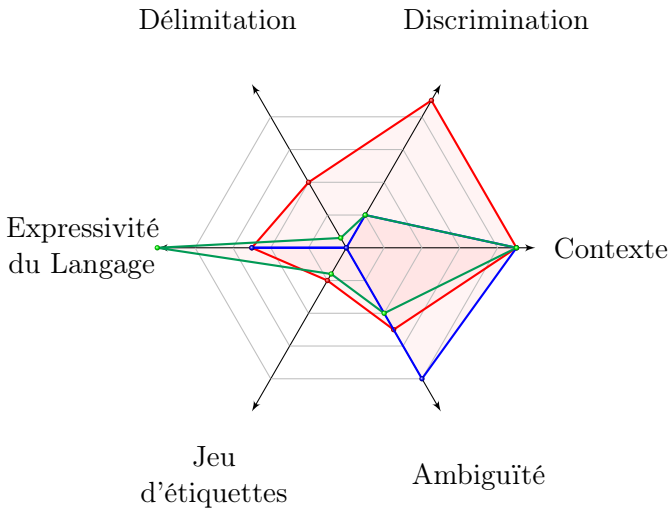


FIGURE 7.6 – Représentation de la complexité de trois campagnes d'annotations selon (Fort 2012) : Discontinuités, Syntaxe et SDRT

Chapitre 8

Extraction d'informations par les marqueurs explicites de discours

L'utilisation des représentations sémantiques et discursives nous a conduit à travailler dans une toute autre direction. Il s'agit de mobiliser les informations discursives pour elles-mêmes dans une autre tâche, en particulier dans l'extraction d'informations. L'idée générale est d'utiliser des représentations linguistiques pour l'inférence de nouvelles connaissances, ce qui fait le lien entre traitement de la langue et des connaissances. Pour cela nous avons collaboré avec Yannick Toussaint, dont la spécialité est l'extraction d'informations. Nous avons co-encadré le travail d'une étudiante de deuxième année de master, Sara van de Moosdijk, (Moosdijk 2014), pour poursuivre, ceux d'un autre étudiant de seconde année de master Anh-Duc Vu (Vu 2016).

Le principe est de travailler sur des articles scientifiques du domaine médical. Nous cherchons des relations de discours, identifiées par des marqueurs explicites, entre des concepts de ce domaine. Ces informations sont ensuite utilisées comme données d'entrées pour des méthodes d'extraction des connaissances comme la FCA (Ganter et Wille 1999) ou des méthodes plus récentes inspirées de la FCA : les *Pattern Structures* (PS) (Ganter et Kuznetsov 2001). Nous souhaitons ainsi identifier de nouvelles connaissances, par exemple entre des maladies et des symptômes.

Actuellement, la grande majorité des approches d'extraction d'informations envisage un documents comme un sac de mots, à partir duquel des mots clés sont choisis et utilisés dans le processus ou des approches basées sur la sémantique distributionnelle. Ces méthodes utilisent des outils pour réaliser des prétraitements (filtres, lemmatisation, catégories morpho-syntaxiques, désambiguïsation), (Hotho, Nürnberger et Paaß 2005) qui fournissent des informations additionnelles sur les mots, mais sans informer sur les relations sémantiques. Pour passer à un niveau plus abstrait, nous avons proposé d'utiliser des éléments extraits de la structure du discours.

Afin d'illustrer le propos, regardons l'exemple 8.1 qui reprend deux phrases extraites d'articles médicaux. La première phrase établit simplement qu'un anesthésique est utile lorsqu'un enfant passe un IRM. Un humain (ayant quelques connaissances) sera capable d'arriver à la même conclusion à partir du second exemple, ce qui sera probablement impossible pour un algorithme tant les énoncés sont différents.

(8.1) a. *Four children underwent magnetic resonance imaging (MRI), which required a general anaesthetic.*

Quatre enfants ont subi une imagerie par résonance magnétique (IRM), qui ont nécessité une anesthésie générale.

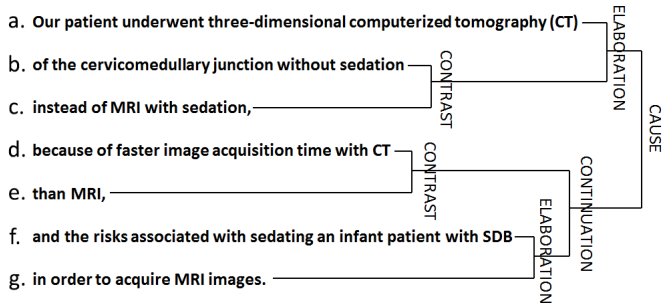


FIGURE 8.1 – Analyse des relations de discours de l'exemple 8.1.b

b. *Our patient underwent three-dimensional computerized tomography (CT) of the cervicomedullary junction without sedation instead of MRI with sedation, because of faster image acquisition time with CT than MRI, and the risks associated with sedating an infant patient with SDB in order to acquire MRI images.*
Notre patient a subi une tomodensitométrie trois dimensions (CT) de la jonction cervicomédullaire sans sédation au lieu de l'IRM avec une sédation, en raison d'un temps d'acquisition des images plus rapide avec CT qu'avec l'IRM, et de moindre risques associés à la sédation d'un enfant avec SDB en vue d'acquérir des images IRM.

Pour modéliser l'exemple 8.1.b, il est possible d'utiliser une représentation proche de la SDRT (Asher et Lascarides 2003), comme dans la figure 8.1. L'analyse du discours aide à structurer les différentes parties de la phrase, et elle permet également d'inférer de nouvelles connaissances. Par exemple, certaines informations que nous pouvons trouver dans l'exemple 8.1.a peuvent aussi être retrouvées dans la relation d'élaboration entre f et g dans la figure 8.1.

Il n'existe que peu de travaux utilisant à la fois les techniques de recherche d'information et des relations discursives. Nous avons donc tenté de montrer l'intérêt de cette approche sur des textes médicaux. Pour faire le lien entre différentes parties de relations de discours, nous utilisons des informations sémantiques spécifiques du domaine (comme des ontologies de termes médicaux). Le résultat de la recherche d'informations pourra seulement être interprété par un expert qui maîtrise ce type de représentation de données.

Le corpus sur lequel Sara von de Moosdijk a travaillé est constitué d'articles extraits de PubMed [1] de manière aléatoire. Elle a choisi douze maladies et pour chacune cinquante articles. Les prétaitements de normalisation et de segmentation ont été réalisés en Python avec `Natural Language Toolkit` (NLTK). Le résultat est un corpus de 600 articles, struturé en XML.

Comme nous l'avons mentionné, nous avons également eu besoin d'informations sémantiques que nous avons ajoutées au texte. Sara von de Moosdijk a utilisé le *Unified*

1. `www.ncbi.nlm.nih.gov/pubmed/`

Medical Language System (UMLS) de l'*U.S. National Library of Medicine*. Cette ressource est composée de plusieurs outils dont un large MetaThesaurus (la version 2014AA contient 2 973 458 concepts). Le lien entre les données et le texte est réalisé avec un reconnaisseur d'entités nommées pour termes médicaux `MetaMap tool` (Aronson 2014).

Il existe plusieurs théories du discours qui tendent à représenter des informations de natures différentes, comme la *Rhetorical Structure Theory* (RST) (Mann et Thompson 1988), *Dynamic Predicate Logic* (DPL) (Groenendijk et Stokhof 1991) ou encore les *Frames* (Barsalou 1992). Nous ne disposons pas pour autant d'outil pratique qui produise les représentations attendues. Certaines recherches d'informations visent à extraire des relations de discours entre quelques phrases, ou d'autres tentent de réaliser la tâche sur des documents de plus grande taille, voire au niveau du document dans son ensemble.

La plupart des outils supposent l'existence de corpus annotés ce qui n'est pas le cas. Par ailleurs, les corpus existants sont développés pour des théories difficilement comparables. (Soricut et Marcu 2003) utilise le *RST Discourse Treebank* (Carlson et al. 2002). (Baldridge et Lascarides 2005) utilisent leurs propres annotations sur des dialogues du *Redwoods Treebank* (Oepen et al. 2002) à partir de la SDRT, alors que (Muller, Afantenos et al. 2012) utilise ANNODIS (Péry-Woodley et al. 2011), un corpus en français pour la SDRT. Enfin, (Wellner et al. 2009) réalise des expériences sur le *Discourse GraphBank* (Wolf et al. 2004). Malheureusement, aucun de ces corpus n'est développé dans le contexte médical.

Il n'est pas possible d'utiliser des outils pour l'annotation en relation de discours dans ce contexte, mais Sara von de Moosdijk a repris (Marcu et Echihabi 2002) et (Sporleder et Lascarides 2008) qui décrivent des marqueurs explicites de relations de discours pour l'anglais. La différence entre les deux est l'ensemble des relations de discours choisies. (Marcu et Echihabi 2002) utilisent : *contrast*, *cause-explanation-evidence*, *condition*, et *elaboration* ; et (Sporleder et Lascarides 2008) : *contrast*, *result*, *summary*, *continuation*, et *explanation*.

Au total Marcu and Echihabi ont listé douze patrons syntaxiques, contenant huit marqueurs de discours distincts ; et (Sporleder et Lascarides 2008) ont utilisé une liste de cinquante-cinq marqueurs de discours. Ces derniers n'ont pas rendu publics leurs patrons syntaxiques, ce qui ne permet pas de reprendre leurs travaux.

La table 8.1a présente l'ensemble final de patrons utilisés pour marquer les relations de discours. L'extraction est réalisée article par article, phrase par phrase. Ici, une phrase ne peut recevoir qu'une seule relation interne, une autre avec la phrase précédente et une dernière avec la suivante (soit trois relations au maximum). L'ensemble des relations de discours utilisées jusqu'à présent est une généralisation faite par Marcu et Echihabi basée sur plusieurs théories du discours, y compris la SDRT. La table 8.1b résume le résultat final de l'identification des relations de discours. Elle donne le nombre de relations de chaque type trouvées dans les textes (soit 82 667 phrases). Les relations restent trop éparses pour obtenir une interprétation significative.

Sara von de Moosdijk utilise la FCA (Ganter et Wille 1999) qui est une méthode de classification et de conceptualisation basée sur des contextes formels (G, M, I) (où G est un ensemble d'objets, M un ensemble d'attributs, et $I \subset G \times M$ une relation binaire

CONTRAST
[BOS ... EOS] [BOS But ... EOS]
[BOS ...] [but ... EOS]
[BOS ...] [although ... EOS]
[BOS Although... ,] [... EOS]
[BOS ... EOS] [BOS However ... EOS]
[BOS Whereas... ,] [... EOS]
[BOS . . .] [whereas ... EOS]
[BOS (In|By) contrast ... ,] [... EOS]
[BOS ... EOS] [BOS (In|By) contrast, ... EOS]
CAUSE-EXPLANATION-EVIDENCE
[BOS ...] [because ... EOS]
[BOS Because ... ,] [... EOS]
[BOS ... EOS] [BOS Thus, ... EOS]
[BOS ... EOS] [BOS Consequently ... EOS]
[BOS ...] [(and)(,) consequently ... EOS]
CONDITION
[BOS If... ,] [... EOS]
[BOS If...] [then ... EOS]
[BOS ...] [if ... EOS]
ELABORATION
[BOS ... EOS] [BOS... for example... EOS]
[BOS...] [which... ,]

(a) Ensemble des patrons syntaxiques utilisés avec BOS : *beginning-of-sentence*, EOS : *end-of-sentence*

Relation	# found	R	Initial R	Marcu's R
Contrast	6545	7.92	5.0	9.43
Cause	1726	2.08	2.05	2.16
Condition	793	0.96	1.17	2.93
Elaboration	4181	5.06	5.24	4.46

(b) Résultats de l'extraction des relations (R signifie « ratio »)

TABLE 8.1 – Patrons syntaxiques par type de relations de discours explicites et répartition dans le corpus

entre G et M).

À partir de ces contextes, il est possible d'appliquer les méthodes de *Pattern Structures* pour construire des représentations qui s'interprètent comme des distances entre les éléments. Pour cela nous appliquons l'algortihme *CloseByOne*. Les *pattern concepts* (Ganter et Kuznetsov 2001) de $(G, (D, \cap), \delta)$ sont des paires de la forme $(A, d), A \subseteq G, d \in (D, \cap)$, où d est la description commune des objets dans A. L'ensemble de tous les *pattern concepts* forme alors un treillis. Des algorithmes existants pour la FCA peuvent être réutilisés avec quelques modifications pour calculer les *pattern structures*.

La structure de treillis que nous pouvons construire à partir de ces relations de discours et les relations sémantiques précédentes donne une image des combinaisons possibles de relations de discours. La combinaison du réseau avec des visualisations des intentions sous la forme de structures d'arbres rend la totalité de la production beaucoup plus facile à interpréter. Dans notre cas, le résultat fait appel à des connaissances extérieures très spécifiques. Le résultat de l'extraction n'est alors accessible qu'aux seuls experts. Il nous faut donc les identifier et organiser une phase d'évaluation qualitative pour pouvoir conclure sur la qualité du résultat obtenu.

Nous ne souhaitons pas revenir sur les détails de cette seconde partie du projet qui n'utilise pas de propriétés linguistiques mais réalise des calculs de similarité pour

construire des ensembles cohérents de documents. Les résultats obtenus sont intéressants, mais n'apparaissent pas suffisamment pertinents pour le moment. La tâche d'évaluation des résultats par des experts manque encore. Nous continuons à travailler à améliorer le processus d'identification des marqueurs de discours, et également celui de mobilisation de ces informations pour leur interprétation.

Chapitre 9

Perspectives et projet de recherche

Sommaire

Ce chapitre présente notre projet de recherche pour les prochaines années. Comme ce document en témoigne, nos recherches portent sur la modélisation de la sémantique des langues naturelles et nous souhaitons poursuivre sur cette thématique. Nous avons abordé les questions relatives au passage de la syntaxe à la sémantique, à l'analyse de phénomènes linguistiques complexes, aux propriétés formelles des formalismes et à leur utilisation dans des contextes réels. La perspective principale pour continuer ces travaux est de proposer des grammaires sémantiques robustes, au sens où elles reconnaissent plusieurs phénomènes en même temps (quantification, événements, présupposition, *etc.*) sur un vocabulaire non restreint. Pour cela, il convient de ramener les aspects théoriques de la modélisation sémantique vers les plus empiriques qui apparaissent dans le projet SLAM. Un axe est de poursuivre sur la définition de modèles sémantiques pour les dialogues qui intègreraient les spécificités des dialogues pathologiques. Il s'agit de mobiliser les théories existantes et la grammaire sémantique précédemment abordée par la modélisation des interactions dialogiques. Cette proposition permet de faire le lien effectif entre la modélisation sémantique de la première partie et le projet SLAM. Une autre perspective est dans le développement de l'outil `SLAMtk`, notamment en cherchant à augmenter le nombre de phénomènes linguistiques analysés. Il sera aussi intéressant de confronter cet outil à des données variées, par exemple des discours et débats politiques. Cette brève présentation met en avant les quatre volets que ce chapitre reprend. Nous revenons tout d'abord sur les évolutions de l'outil `SLAMtk` dans la section 9.1, puis sur les évolutions de SLAM pour la modélisation plus fine des entretiens pathologiques dans la section 9.2. Nous discutons ensuite des grammaires à large couverture dans la section 9.3 et enfin la modélisation des dialogues formels dans la section 9.4 avant de conclure ce document.

9.1 Évolutions de SLAMtk

La part de nos recherches la moins directement portée sur la sémantique est celle qui concerne non pas le projet SLAM, sur lequel nous reviendrons par la suite, mais

sur le développement de l'outil `SLAMtk`. Comme nous l'avons abordé, cet outil est déjà opérationnel et génère automatiquement des annotations en disfluences et en catégories morpho-syntaxiques. Pour chaque jeu d'étiquettes proposé, l'outil calcule des différences de répartition entre des phénomènes pour mettre en avant des usages spécifiques.

9.1.1 Analyse de la syntaxe

Augmenter la couverture linguistique de l'outil est une question importante. Nous avons envisagé à ce titre de produire des représentations syntaxiques et de les mobiliser pour l'analyse. Comme la production langagière des schizophrènes semble standard pour la morpho-syntaxe, il serait intéressant d'analyser ce qu'il en est pour la maîtrise des structures syntaxiques complexes. On pourrait par exemple étudier les enchâssements multiples de relatives qui sont considérés comme des indicateurs de compétence cognitive.

Pour cela, il nous faudrait utiliser un analyseur syntaxique développé pour le français, à partir des résultats duquel on extrairait ces informations spécifiques. Bien que le principe puisse paraître simple, sa mise en œuvre n'est pas évidente, en particulier parce que nous travaillons à partir de transcriptions de l'oral où la notion de phrase avec structures syntaxiques bien définies n'y est pas triviale. Par exemple, (Deulofeu et al. 2010) argumentent pour une refonte en profondeur du schéma d'analyse de la syntaxe de l'oral. Contrairement à eux, on trouve des propositions comme (Anne Abeillé et B. Crabbé 2013) qui tentent de transférer les structures d'annotations de la syntaxe de l'écrit vers celles de l'oral. Ils proposent une notion de phrase non-triviale. On peut choisir de considérer des phrases phonétiques à partir des pauses dans la conversation, des phrases dialogiques qui correspondent à un tour de parole, des phrases discursives qui correspondent à un acte de langage, et une notion syntaxique qui correspond à une plus grande unité syntaxique complète (avec enchâssement possible). On pourrait alors utiliser plusieurs éléments de ces définitions pour parvenir à un modèle général. Un tour de parole introduit des phrases, et chaque tour de parole peut être redécoupé en sous-phrases. Dans ce cas, un acte de langage sur plusieurs tours de parole ne construit pas de phrase unique, et il est préférable de travailler sur des phrases inachevées.

La démarche est confirmée par le nombre de ressources et outils disponibles pour l'écrit car en disposer pour le traitement de l'oral serait une plus-value importante. Il ne faut cependant pas considérer que les outils puissent légitimer à eux-seuls une théorie. Afin de reconstruire des unités syntaxiquement plus proches de l'écrit, on pourrait identifier les disfluences. Or ces dernières font déjà partie de notre schéma d'annotations et il est aisé de produire une version du corpus qui ne les contient pas. Bien que cette proposition puisse apparaître comme minimale, (Anne Abeillé et B. Crabbé 2013) ont comparé (1) l'application d'outils d'analyses syntaxiques sur des données de l'oral et des données pré-traitées (sans les disfluences), (2) des outils entraînés sur des données de l'oral et (3) une méthode combinant les deux approches. Il apparaît que le corpus sans disfluence conduit aux meilleurs résultats avec une précision de 76%.

Pour tester ces approches, nous avons utilisé différents analyseurs disponibles sur le corpus. En particulier, nous avons d'ores et déjà procédé à une analyse syntaxique automatique classique grâce aux outils `Talismane` (Urieli et Tanguy 2013) et `FRMG` (De

La Clergerie et al. 2009). Les données n'ont pas encore été analysées en profondeur. Cependant, une première constatation sur les résultats ne fait pas apparaître de complexité syntaxique particulière qui semble équivalente chez les schizophrènes par rapport au groupe témoin. Au delà de ces premières lectures, il n'est pas aisé de proposer en l'état une analyse convaincante de l'interprétation de la complexité syntaxique par rapport à la qualité des résultats obtenus. Il convient donc de poursuivre dans cette voie.

9.1.2 Production automatique de la transcription

Comme nous l'avons indiqué précédemment, nous ne pouvons pas travailler à partir d'une version dégradée de la ressource, la qualité de la transcription est donc un enjeu important. Cependant, le projet est très lourd à porter car la gestion de la transcription manuelle demande beaucoup de temps. Il faut d'une part obtenir les ressources financières pour produire ces transcriptions et d'autre part contrôler la régularité de la transcription tout au long du processus.

Plusieurs tests ont été menés en 2013 pour utiliser des outils existants pour la transcription automatique de l'oral. Ce type de système est généralement construit à partir d'enregistrements d'émissions de radio. Nos données sont assez proches de ces contraintes, par exemple elles contiennent peu de recouvrements entre les intervenants.

Le test le plus poussé a été réalisé avec le logiciel JTRANS, (Cerisara, Mella et Fohr 2009) qui apparaissait comme de bon niveau pour la manipulation de données en français. Les résultats se sont révélés décevants car moins de 10% des mots étaient correctement reconnus, et les tests de temps de correction se sont avérés trop peu efficaces pour que l'outil soit réellement utilisé.

Depuis, de nombreux résultats sur la question du traitement de l'oral, basés sur plusieurs méthodes numériques autour de la modélisation acoustique ont été présentés. Bien que des limites existent pour arriver à des performances qualitativement acceptables, des solutions émergent, par exemple dans la prise en compte des caractéristiques techniques de la prise de son par des modèles de Markov cachés (HMM) dynamiques. Une autre perspective est l'utilisation de réseaux de neurones profonds (DNN) (Abdelaziz et al. 2015 ; Orosanu et Jouvet 2015) qui permettent de rendre les systèmes plus précis et plus robustes.

Des tests sont actuellement en cours, avec les nouvelles versions de JTRANS et les ressources développées dans l'équipe Multispeech. Les évolutions techniques et technologiques laissent penser que la qualité des résultats pourrait être suffisamment raisonnable pour être utilisées. Nous envisageons dans ce cas de mettre en place une évaluation du temps de correction pour redresser les résultats. Il s'agirait de compter le nombre de corrections, ainsi que le temps de correction à apporter dans un outil d'annotation par plusieurs annotateurs qui travailleraient sur des résultats provenant de différents outils et de différentes ressources.

9.1.3 Autres évolutions

De manière plus générale, l'outil `SLAMtk` doit connaître plusieurs évolutions pour être efficace dans l'analyse automatique. Un premier aspect consiste à intégrer de plus nombreux tests. On retrouve des inspirations dans la statistique lexicale comme le *Burrows' Delta*, le *Chi-square*, le *Kullback-Leibler Divergence*, le *Z-score* ou la distance intertextuelle qui se définit par la différence entre les productions de deux intervenants.

$$D(A_1, A_2) = 1 - \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|}$$

Il est aisé de calculer ces différences entre chaque interlocuteur et ainsi de chercher si des ensembles de comportements émergent. Ces données peuvent être calculées par exemple sur la répartition des mots, des lemmes ou des catégories morpho-syntaxiques. Il serait également intéressant d'analyser ces résultats en fonction de la définition de l'unité linguistique choisie comme nous en avons discuté précédemment. Est-ce que ces résultats restent en moyenne valides en fonction de leur apparition dans ces unités ?

Par extension, il apparaît judicieux de mettre en place des analyses multivariables comme l'analyse en composantes principales (PCA) ou des analyses de la variance. Ces méthodes sont avérées et disponibles dans des bibliothèques *ad hoc*. Il s'agit plus particulièrement de les intégrer correctement dans l'environnement pour extraire automatiquement des indices pertinents. Toutes ces évolutions sont des intégrations directes à apporter à l'outil et à mobiliser pour l'analyse. Il semble évident que l'utilisation de `SLAMtk` dans des contextes différents, par exemple pour l'analyse de discours politiques, fera naître des besoins spécifiques pour ces analyses.

Une partie plus complexe réside dans l'analyse lexicographique. Si nous avons pu proposer des analyses manuelles dans (Amblard, Fort, Musiol et al. 2014) à partir de l'outil `TXM` (Heiden 2010), l'automatisation paraît plus difficile à mettre en œuvre. Certains prétraitements peuvent être définis - comme l'identification des formes les plus fréquentes - en supprimant les unités sémantiquement vides des ensembles de formes traitées. Des vecteurs sémantiques, comme ceux utilisés dans la sémantique distributionnelle (Harris 1954) pourraient être employés pour proposer des usages caractéristiques dans chaque entretien et pour chaque type d'interlocuteur.

Il s'avère que l'outil n'est pas encore distribuable largement et simplement car l'implémentation de ses fonctionnalités reste trop spécifique à son utilisation originelle. Un premier travail de normalisation autour d'une structure de représentation intermédiaire de type XML est en cours de finalisation. Il s'agit de partir d'une version écrite des entretiens dont la transcription suit le guide prédéfini. L'outil extrait les données nécessaires à l'analyse et produit une représentation XML augmentée des annotations automatiques. Il est ensuite possible d'utiliser cette représentation pour générer les analyses dont nous disposons déjà, et aussi d'ajouter celles que nous venons de décrire. Cette version normalisée pourra aisément être utilisée sur d'autres corpus, ainsi que sur des données d'autres types.

Il serait nécessaire d'implémenter une interface graphique pour simplifier la compréhension de l'outil. Nous proposons de nombreux traitements, mais il est difficile d'utiliser

toutes les fonctionnalité de l'outil sans les connaître à l'avance. D'autre part, il est aussi possible de produire des représentations dont l'interprétation n'est pas évidente. Il paraît important d'accompagner l'utilisateur dans sa lecture des résultats plutôt que de lui délivrer des résultats bruts.

Enfin, comme nous l'avons évoqué précédemment, l'analyse syntaxique mériterait d'être affinée. Par ailleurs, la poursuite sur l'analyse des niveaux phonétiques et phonologiques serait une piste intéressante.

9.2 Modélisation des entretiens des patients schizophrènes

Pour poursuivre le projet SLAM, nous avons plusieurs perspectives selon des plans opératoires de natures différentes : d'un côté travailler sur la constitution d'un nouveau corpus, et d'autre part travailler à la modélisation des interactions.

9.2.1 Constitution d'un corpus augmentant les pathologies étudiées

Comme nous l'avons exposé, le projet SLAM a donné des résultats encourageants dans cette première phase concernant l'identification de symptômes spécifiques, à partir de l'analyse de différents niveaux linguistiques. Cependant, plusieurs conjectures demeurent sur le fonctionnement cognitif ou la complexité linguistique des phénomènes, et il nous semble nécessaire de poursuivre dans cette direction.

Pour cela, nous souhaitons augmenter et enrichir le corpus de données empiriques en procédant à de nouveaux entretiens avec des patients schizophrènes, et en ouvrant à d'autres pathologies autour des troubles de la pensée. À partir de ces nouvelles données, nous projetons d'approfondir la méthodologie de construction des représentations que nous définissons depuis plusieurs années.

Une nouvelle phase de recueil de données est en cours en partenariat avec le centre Hospitalier Montperrin à Aix-en-Provence. Nous disposons à ce jour d'une autorisation d'intervention. Le centre comporte une unité médicale appropriée et un personnel thérapeutique est dégagé pour participer à l'étude. Il s'agit de rencontrer soixante patients (paranoïaques et schizophrènes, en ambulatoire ou en programme de soins). De manière symétrique, nous prévoyons de constituer un groupe témoin apparié de soixante personnes. Ce programme est très ambitieux eu égard au nombre de patients envisagés, mais notre méthodologie étant bien définie, ces objectifs semblent réalistes. Nous disposons déjà de 7 entretiens.

Le dernier corpus que nous avons étudié a été constitué à partir d'une population en remédiation avancée. L'apparition marginale de discontinuités n'est donc pas en soit un résultat négatif, mais nous fournit un argument sur la remédiation (qui permettrait de faire disparaître certains symptômes). Il convient donc de confronter la modélisation à des données plus proches de l'expression de la pathologie. Cette nouvelle phase de l'étude a une double fonction. Il s'agit d'une part d'éprouver nos hypothèses de dysfonctionnement sémantico-pragmatique auprès de patients moins avancés dans la remédiation. D'autre part, elle nous permet d'augmenter le nombre de manifestations de discontinuités et

donc de réfléchir et d'argumenter sur la nature des processus cognitifs mobilisés. Si les phénomènes qui nous intéressent sont typiques, ils n'en demeurent pas moins rares. Le travail sur la pathologie des schizophrènes n'est donc pas terminé.

Par ailleurs, il paraît nécessaire d'interroger les pratiques linguistiques pour des pathologies proches de la schizophrénie. Par exemple, la schizophrénie partage une partie de ses symptômes de troubles de la pensée avec d'autres pathologies. Il est donc important de tester si ces symptômes langagiers apparaissent ou non pour ces maladies connexes. Une première série de travaux avait été proposée pour recueillir des entretiens avec des patients atteints d'autisme de haut niveau (type syndrome d'Asperger), mais pour plusieurs raisons techniques, nous n'avons pas été en mesure de collecter ces données. Il avait cependant été nécessaire de réfléchir aux traitements potentiellement mobilisables sur ce type d'entretien dont la nature resterait *de facto* éloignée de l'interaction dialogique traditionnelle.

Ce type d'argumentation est donc à reprendre, à intégrer à notre réflexion et à éprouver auprès de ces autres pathologies. Comme la sous-catégorie de schizophrènes particulièrement concernée par ces dysfonctionnements sémantico-pragmatiques est la catégorie des schizophrènes paranoïdes, il nous paraît important de nous confronter à la paranoïa. Cette première étape n'est évidemment pas exclusive de l'étude des autres formes de schizophrénie, de schizotypie [1] ou de dépressions.

9.2.2 Modélisation des interactions et des processus cognitifs

La seconde perspective sur la modélisation est de déployer les outils techniques, méthodologiques et théoriques dont nous disposons pour analyser ce corpus. Le cœur de notre étude reste l'analyse sémantico-pragmatique. Les éléments de formalisation précédemment proposés pourront ainsi être remis en question et affinés.

Ce travail de modélisation à partir de la SDRT nous a permis d'expliciter les enjeux de notre recherche. Cependant, nous avons mis en évidence plusieurs limites sur lesquelles nous devons poursuivre notre réflexion. On se demande naturellement quelle est la position de la modélisation du savoir partagé entre les interlocuteurs. On peut supposer, dans le prolongement de Clark (H. Clark et E. Clark 1977 ; H. Clark, Schreuder et Buttrick 1983), que les interlocuteurs disposent d'un savoir mutualisé (*common ground*) qui contient les informations nécessaires à la réalisation des inférences ou des implicatures pour parvenir à l'interprétation des interactions complexes. Mais, nous n'avons pas de définition, ni de modélisation des processus permettant de mobiliser les informations du savoir mutualisé. Travailler sur ces questions permet d'étudier les opérations cognitives, pour la pathologie mentale ou pas. Aborder le syndrome chez les schizophrènes nous permet de continuer à travailler sur le fonctionnement cognitif en général.

Il semble par ailleurs que nous puissions identifier une convergence entre les résultats sur le lieu du dysfonctionnement. Par exemple, nous n'identifions pas de difficulté au niveau morpho-syntaxique, mais d'autres résultats mettent en avant une diminution de

1. La schizotypie est une maladie possédant des troubles de la pensée dans sa symptomatologie, comme la schizophrénie, mais dont les implications dans la maladie sont moindres que celles de la schizophrénie. Par exemple, les patients ne souffrent pas d'hallucinations.

la complexité syntaxique [2], ou encore nous voyons apparaître des comportements particuliers pour la disfluence sans apparition de trouble sur la segmentation. La base neurocognitive du langage ne semble pas porteuse de défaillance particulière. La nécessité de mobiliser la compétence dans un contexte engendre les dysfonctionnements.

On peut supposer que l'incapacité des schizophrènes à maintenir la cohérence de leur discours pourrait venir de leur incapacité à mobiliser ce savoir mutualisé. Dans ce cas, certaines inférences qui auraient permis de poursuivre l'interaction n'ont pas lieu et le dialogue doit subir des adaptations pour rester interprétable. Ces inférences sont le résultat des opérations cognitives que nous avons précédemment abordées. Elles trouvent ainsi un support et une fonction qui doivent être intégrés à la formalisation pour expliquer (et expliciter) leur rôle dans la construction de la cohérence.

La prise en compte de ces éléments doit s'accompagner de modifications profondes dans la représentation. L'utilisation d'extensions récentes de la DRT, voire de nouvelles représentations, sont nécessaires. Des pistes sont à explorer à partir de (von Heusinger et Meulen 2013) qui intègrent la projection des croyances d'un individu pour la construction de sa représentation sémantique. Ce principe se rapproche de celui que nous présupposons pour expliquer les incohérences apparaissant dans les interactions. Dans notre cas, il s'agirait de projections erronées de croyance ou de connaissance sur la représentation de l'interlocuteur. La non-compréhension provient alors de la différence entre les inférences faites, ou supposées possibles, à partir des représentations singulières de chacun des locuteurs.

Il n'est malheureusement pas possible de proposer une théorie et ses contreparties formelles aussi directement. Il faut en effet revenir sur la notion de connaissance et définir le calcul compositionnel permettant d'extraire les informations pertinentes. Mais ces éléments bousculent la notion de propositions et par la même celle de contexte. La SDRT d'origine doit être repensée pour formuler une représentation cohérente.

Nous nous intéressons à ces questions, sans omettre que nos données sont issues de dialogues et non de discours. En cela, la SDRT n'est pas la formalisation la plus adaptée, mais ces phénomènes sont plus explicitement la marque de processus cognitifs en action et ne peuvent donc pas apparaître dans des textes écrits. La formalisation doit cependant les intégrer. Nous reviendrons sur cette question dans la section 9.4, où l'une des ouvertures est la modélisation formelle des dialogues.

Nous n'avons pas encore mentionné dans ce chapitre que la formalisation dans SLAM est basée sur une extension de la SDRT et non de la DRT. Or, le contexte de notre travail pousse à ajouter à cette refonte les notions de contexte et de continuation du calcul (de Groote 2006) qui apparaît d'autant plus adapté que la modélisation de processus cognitifs doit s'inscrire dans une vision procédurale du calcul sémantico-pragmatique et pas seulement représentationnelle. Il faut bien distinguer ici la modélisation des dialogues en général, de la mobilisation de la formalisation pour des interactions pathologiques. Ces deux questions appartiennent bien à la même perspective de recherche, sans se recouvrir (modélisation de processus cognitifs et capacité de représentation des dialogues).

2. Nous pensons pouvoir confirmer ce type de résultats avec nos extensions de `SLAMtk` avec des analyseurs syntaxiques.

Il ne faudrait pas croire que nous nous inscrivons radicalement en faux des propositions (Kamp et Reyle 1993), mais bien que nous essayons de réconcilier ses propositions avec nos modélisations. Hans Kamp a toujours basé son argumentation sur l'existence de représentations mentales pour justifier des représentations sémantiques proposées, sans intégrer la psychologie à ses représentations. Ici, nous tentons donc d'inclure des arguments motivés par la psychologie sans remettre en question son approche.

Enfin, revenir sur la définition du cadre d'interprétation implique de revenir sur la définition des contextes d'interprétation. Un travail est actuellement en cours pour proposer une typologie des contextes qui contient (Rebuschi 2015) :

— le contexte discursif qui dépend de l'interaction et de la dynamique d'interaction ;
— le contexte doxatique qui reprend l'ensemble des présuppositions, des croyances à propos du monde et la projection des croyances des locuteurs ;
— le contexte pragmatique qui s'interprète par la situation de l'interaction (le locuteur qui dit « je » en jouant un rôle ne dit pas « moi » pour se désigner lui, mais pour désigner l'individu qu'il cherche à incarner) ;
— le contexte matériel et social où l'idée est de considérer à la fois le cadre de l'interaction, jusqu'à l'ensemble des influences qui la construisent.

À partir de cette première proposition de classification des éléments qui modifient l'interprétation de l'interaction, on peut situer le savoir mutualisé que nous venons d'introduire. La modélisation formelle ne peut évidemment pas intégrer tous ces contextes directement, mais définir des représentations selon chacune de ces dimensions le plus fidèlement possible.

Le projet SLAM a encore plusieurs axes sur lesquels se déployer, tant du point de vue pratique sur la constitution de ressources que théorique pour la modélisation de l'interaction.

9.3 Grammaire sémantique à large couverture

Actuellement, la sémantique et les applications sémantiques sont majoritairement traitées par des approches numériques qui s'intéressent à construire des univers représentant les relations existantes dans les énoncés plutôt qu'à leurs aspects vériconditionnels. On appelle ce type d'approche la sémantique distributionnelle (Harris 1954 ; S. Dumais 2003). Les formalisations les plus utilisées sont la *Latent Semantics Analysis*, (Landauer et Dutnais 1997 ; S. T. Dumais 2004) ou *Vector Space Models* (Erk 2012 ; S. Clark 2015). Pour ces approches, il reste ainsi plus difficile de traiter de phénomènes sémantiques particuliers car elles ne disposent pas d'une représentation structurelle de la sémantique.

L'approche montagovienne de la sémantique permet d'obtenir ce type de représentation, mais, actuellement, il n'existe pas de grammaire à large couverture réaliste pour le français. S'il est possible de proposer le traitement de phénomènes particuliers (négation, intensionalisation, *etc.*), la question de la combinaison de ces traitements est en elle-même une autre difficulté. Plusieurs théories de la modélisation du discours sont installées dans le paysage scientifique, mais peu de ressources existent. On retrouve en particulier le *Groningen Meaning Bank* (Bos et al. 2016) qui fournit des représentations

en SDRT pour l'anglais.

Disposer de ces représentations permettrait aux systèmes utilisant la compréhension fine des énoncés en langue naturelle de faire un saut qualitatif. Les applications sont nombreuses du côté des systèmes de dialogue, des interactions Homme-Machine, ainsi que l'identification des éléments pertinents dans de grandes masses de données.

9.3.1 Constitution de la grammaire

Comme nous l'avons vu dans les chapitres précédents, nous disposons d'un environnement permettant de développer des grammaires sémantiques pour reconnaître des phénomènes particuliers. Nous travaillons également à disposer d'un cadre dans lequel il serait possible de développer des fragments spécifiques qui pourraient être composés, mais nous sommes confrontés à une question récurrente autour de l'évaluation. Une partie du problème est de disposer de lexique suffisamment grand pour procéder à des analyses de données non contrôlées (traitement d'exemples en dehors de grammaires jouets).

Les ACG (de Groote 2001) sont un cadre grammatical dans lequel de nombreux formalismes grammaticaux peuvent être encodés. Elles offrent un contrôle aussi bien sur les structures d'analyse que sur les structures de surface (par exemple une suite de mots). Elles proposent des traitements unifiés permettant de dériver simultanément plusieurs représentations linguistiques : en suite de mots, en structures syntaxiques, en formules logiques, *etc.* Elles sont basées sur le partage de structures abstraites qui permettent de modéliser l'interface entre la syntaxe et la sémantique. Si le cadre est bien étudié, sa mise en œuvre sur des exemples réalistes n'est pas aboutie. L'outil développé autour de la théorie, `ACGtk`, est encore uniquement utilisé sur des petites grammaires. Les ACG sont donc le cadre adapté pour lequel il faut développer des ressources spécifiques.

Les grammaires d'arbres adjoints (TAG) sont un formalisme grammatical très répandu pour l'analyse des langues naturelles. Plusieurs grammaires à large couverture ont été développées, en particulier pour le français. En plus de la modélisation syntaxique, SemFrag (Gardent et Parmentier 2005 ; Gardent 2008) permet la construction de représentations sémantiques au moyen de formules logiques à partir des analyses syntaxiques. Le lexique Frigram (Perrier et Guillaume 2013a) qui est développé pour l'analyseur Léopar et les grammaires d'interaction (Perrier et Guillaume 2013b), est un autre exemple d'une telle ressource pour le français.

Ces deux ressources nous intéressent car elles proposent une couverture du français de bonne qualité. Par ailleurs, les TAG peuvent être encodées dans les ACG (Pogodalla 2004 ; Pogodalla 2009), ce qui nous assure de pouvoir transférer le lexique de SemFrag vers notre nouvelle grammaire. Dans le même temps, la proposition faite pour la modélisation sémantique dans SemFrag se heurte à des difficultés, notamment dues à la gestion de la compositionnalité et plus encore à la gestion de la notion de contexte. Un simple transfert de la modélisation des continuations ne semble pas suffire pour obtenir la couverture sémantique attendue.

Par ailleurs, les hypothèses théoriques qui fondent Frigram sont communes aux ACG. Les deux formalismes sont développés à partir de la logique linéaire. La ressource a cependant été construite pour gérer la syntaxe, et des choix spécifiques ont été formulés. Cette

spécialisation explique qu'il ne soit pas aisé de transférer directement la grammaire de la syntaxe à la sémantique, en particulier parce que l'encodage des grammaires d'interaction dans les ACG nécessite des adaptations pour la gestion des types complexes.

Dans les deux cas, nous disposons de ressources dont la couverture pour le français est indiscutablement importante et nous pouvons proposer des perspectives explicites sur la méthodologie à suivre pour construire la ressource sémantique dans les ACG. Reprendre les principes des continuations dans le λ-calcul par ces formalismes risque d'être compliqué à cause de la définition de mécanismes *ad hoc*, alors que les ACG peuvent être appréhendées comme un cadre unifiant des différents apports.

L'objectif est double : utiliser la grammaire de SemFrag pour construire une grammaire sémantique à large couverture dans le cadre ACG, et expérimenter cette grammaire avec l'outil de développement des ACG, `ACGtk`. Ces tests permettront d'identifier les besoins pour la couverture de la grammaire qui pourront être comblés par l'intégration d'informations provenant de Frigram.

Des études préliminaires ont pu être réalisées par des étudiants encadrés lors de leur licence et par des étudiants co-encadrés, avec Sylvain Pogodalla, lors de leur seconde année de Télécom-Nancy. Ces études ont montré que la transformation n'était pas *ad hoc* et qu'il convenait de travailler à l'interprétation des propriétés utilisées dans chacune des grammaires pour leur attribuer une contrepartie sémantique. Plusieurs tâches apparaissent :

- la transformation est majoritairement réalisable rapidement. Il convient de la tester sur des exemples issus de corpus réels ;
- des questions spécifiques comme l'interprétation des structures de traits qui contrôlent largement la surgénération des formalismes restent ouvertes ;
- les résultats doivent être étendus à des phénomènes linguistiques particuliers : résolution des anaphores pronominales, chaînes de coréférences, présupposition (Beaver 1997), pluriel, temps et aspects, et modalité.

Plusieurs éléments de réponse ont été apportés dans ce document, mais ils doivent être confrontés à l'automatisation.

9.3.2 Tests de couverture des grammaires sémantiques

Une fois les développements de la grammaire sémantique réalisés, il conviendra de la tester. Une première solution serait d'utiliser les mêmes corpus que ceux proposés pour l'évaluation de (Gardent 2008). Le premier avantage est que si le corpus en question est relativement limité il reprend de nombreux phénomènes sémantiques complexes mais déjà identifiés. Cette étape permettrait de construire une grammaire couvrant de nombreux phénomènes, sans toutefois être trop volumineuse. Il ne faut pas négliger que la complexité des algorithmes dépend de la taille des données d'entrée. Ainsi, travailler sur cette version permettrait de faire plusieurs tests qualitatifs.

Il conviendrait par exemple de baser les premières étapes sur la construction de représentations logiques à la Montague, comme (Dowty 1979). Il faut par exemple intégrer ensuite à ces versions la modélisation des événements dans des versions néo-davidsoniennes, et/ou des éléments de temps et aspects des verbes, voire de temporalité en général.

Une autre perspective pour l'évaluation est d'utiliser le *Groningen Meaning Bank* (Bos et al. 2016). Cette ressource est développée pour l'anglais et propose des représentations en DRT. La chaîne de traitement est basée sur un analyseur syntaxique CCG et `Boxer` (Bos 2008) pour la sémantique. L'outil est développé depuis plusieurs années et la couverture est très bonne. Il est donc possible d'analyser des énoncés en anglais avec ces outils et les traductions en français avec les nôtres. La représentation sémantique ne dépendant pas de la langue, les deux résultats doivent être proches l'un de l'autre.

Avec le même objectif, mais pour une tâche un peu différente, il est possible d'utiliser les ressources proposées pour l'évaluation de SemEval, par exemple la *task 1* sur la *Semantic Textual Similarity*. Il s'agit de proposer des représentations pour deux énoncés et de définir le degré d'équivalence entre les deux. S'il peut paraître possible de réaliser cette tâche, il ne faut pas négliger que la plupart des participants utilisent des méthodes numériques avec de très bonnes performances. Cependant, la ressource étant en anglais, il nous faudrait nous inscrire dans la même démarche que dans la proposition précédente en passant par des traductions.

En plus de la partie sémantique, ce projet ne peut pas être complet sans un travail au niveau du discours et des dialogues. Pour cela, la chaîne de traitement doit être testée pour des modélisations inspirées de SDRT sur les relations de discours. Des travaux récents proposent des lexiques de connecteurs de discours explicites et seront un point de départ, tout en ouvrant la question aux relations de discours explicites (Roze, Danlos et Muller 2012) pour le *French Discourse Tree Bank* (Danlos et al. 2012). Des expérimentations proches ont été réalisées dans le projet Annodis (Afantenos et al. 2012) avec des annotations en relation de discours pour des ressources en français (Est Républicain, Wikipédia, Actes du Congrès Mondial de Linguistique Française 2008, Rapports de l'Institut Français de Relations Internationales). Nous pouvons utiliser ces données comme étalon pour l'évaluation des représentations plus élaborées.

Le but de ces tests est de parvenir à disposer d'un outil suffisamment performant pour être utilisé dans des contextes non contrôlés et nous l'utiliserons certainement dans le cadre du projet SLAM. Les corpus de ce projet peuvent donc être largement repris pour l'évaluation des outils. Dans ce cas, les outils devront être particulièrement robustes.

9.4 Formalisme logique pour la modélisation des dialogues

Comme nous l'avons évoqué, la définition d'un formalisme rendant compte des interactions sous forme de dialogue est une perspective importante. Nous avons co-encadré avec Jirka Maršík le travail de master de Stéphan Tiv sur ce thème (Tiv 2016).

Un discours et un dialogue diffèrent en ce qu'une interaction se met en place entre plusieurs points de vue dans un dialogue (Penstein Rosé et al. 1995 ; Lascarides et Asher 2008). En d'autres termes un discours se construit linéairement comme un tout cohérent, alors qu'un dialogue ne devient cohérent qu'après la résolution des ambiguïtés qui apparaissent entre les intervenants. Il ne s'agit pas ici de calculer la sémantique de chaque énoncé pour l'insérer dans un contexte, mais de définir la structure qui permet, à partir de la sémantique d'un énoncé, de négocier par l'interaction jusqu'à obtenir un accord

entre les participants, et alors seulement cette version négociée est insérée dans la représentation générale.

Nous devons ensuite considérer à nouveau la notion de contexte pour intégrer une structure intermédiaire où l'interaction prend son sens en ce qu'elle permet d'interpréter dynamiquement des représentations potentiellement incohérentes. Pour prendre en compte ce type de situation, nous sommes revenus sur les modélisations des dialogues en DRT afin d'en rendre compte dans notre cadre. Nous retrouvons des exemples pertinents dans (Muller et Prévot 2008) dont celui de l'exemple 9.1.

(9.1) A_1 Tu tournes à gauche juste avant Monoprix.
B_2 La rue à sens unique ?
A_3 Oui, celle-là.
A'$_3$ Non, celle d'après.

Une représentation néo-davidsonienne simplifiée de A_1 est : $\exists es.agent(e, tu) \wedge turn(e) \wedge ontheleft(e) \wedge street(s) \wedge before(s, monoprix) \wedge loc(e, s)$. Ici nous avons un événement « tourner », réalisé par « *tu* », dont la localisation s est avant le Monoprix.

Le second interlocuteur poursuit en B_2 avec une question de confirmation qui inclut une propriété supplémentaire. On peut schématiquement représenter son intervention par :

$$\lambda Pes.\exists s'.P(es) \wedge street(s') \wedge oneway(s') \wedge loc(e, s') \wedge s = s'$$

où s' est une nouvelle rue qui intervient dans un événement e qui doit être défini par la propriété P.

Nous cherchons à déterminer une variable s' qui est une rue à sens unique et qui doit s'unifier avec une autre variable. Cette représentation devait être basée sur une représentation plus structurée comme proposé par les *frames* (Barsalou 1992). Cette discussion ne prenant en compte que les grands principes, nous ne revenons pas sur la modélisation de la question. Il faut cependant noter que c'est un élément structurant à considérer sérieusement, car une question attend une réponse ce qui participe exactement au niveau de négociation que nous avons abordé. Et, comme nous l'avons vu dans la section 6 sur les dialogues pathologiques, la présence ou non de cette réponse est utile à l'interprétation.

Si l'interlocuteur obtient une confirmation comme dans A_3, le calcul revient à la construction d'une représentation qui est la composition des deux interventions. Le résultat produit une formule du type :

$$\exists ess'.agent(e, tu) \wedge turn(e) \wedge ontheleft(e) \wedge street(s) \wedge before(s, monoprix) \wedge loc(e, s) \\ \wedge street(s') \wedge onway(s') \wedge loc(e, s') \wedge s = s'$$

Le point important dans cette représentation est l'unification des deux variables qui représentent les routes. On peut simplifier en utilisant une formule équivalente :

$$\exists es.agent(e, tu) \wedge turn(e) \wedge ontheleft(e) \wedge street(s) \wedge before(s, monoprix) \wedge loc(e, s) \\ \wedge oneway(s)$$

Dans cet exemple, on voit d'ores et déjà que la notion de contexte dynamique est nécessaire pour reconstruire la représentation. L'utilisation directe de TTDL n'est cependant pas envisageable car la cohérence ne peut pas être garantie tout le long du processus calculatoire. En effet, si la réponse à la question est A'$_3$ la représentation contient des informations contradictoires. L'interprétation prend une toute autre forme.

$$\exists ess's''.agent(e,tu) \wedge turn(e) \wedge ontheleft(e) \wedge street(s) \wedge before(s, monoprix) \wedge loc(e,s) \wedge street(s') \wedge oneway(s') \wedge \neg loc(e,s') \wedge s \neq s' \wedge street(s'') \wedge loc(e,s'') \wedge after(s',s'') \wedge s = s''$$

Dans cette nouvelle représentation nous conservons les variables et les descriptions des deux rues, et nous spécifions que la seconde rue n'est pas la localisation de l'événement. Par ailleurs, nous pouvons inférer qu'il existe une autre rue qui elle est celle de la localisation de l'événement. On observe ici les processus cognitifs dont nous avons parlé dans la partie 9.2.2. Dans le même temps, nous n'avons pas abordé la question traitée à l'origine dans (Muller et Prévot 2008) sur la nature des rattachements entre chacun des tours de parole.

Ces relations peuvent avoir une influence structurelle importante sur le résultat produit. C'est typiquement le cas lorsqu'un interlocuteur reprend et modifie ce qui vient d'être énoncé. L'exemple 9.2 donne un tel exemple qui est d'autant plus complexe à analyser qu'il n'est composé que de quelques mots. Seules les relations entre les tours de parole construisent le sens.

(9.2) A_1 : Réunion demain.
B_2 : Lundi ?
A_3 : Demain.

La première phrase est utilisée ici pour poser les éléments factuels de la représentation. Il s'agit d'introduire la réunion et le temps de cette dernière.

$$\exists e.meet(e) \wedge date(e, tomorrow) \tag{9.3}$$

La prise de parole B_2 n'est composée que d'un seul mot, mais ce simple mot signifie que l'interlocuteur reprend à sa charge la représentation précédente, mais pas tout le contexte, uniquement la dernière assertion, et modifie l'un des éléments. Comme le mot désigne une unité de temps, la modification porte sur le seul prédicat temporel. Il s'agit d'opérer sur un trait spécifique de la représentation. Il nous faut donc disposer d'un mécanisme extra logique nous permettant de revenir sur la modélisation précédente. La représentation simplifiée de cette forme interrogative est présentée dans la formule suivante :

$$\lambda Pe.\exists e'.P(e) \wedge meet(e') \wedge date(e', monday) \wedge e = e'$$

La composition des deux représentations produit la formule suivante :

$$\exists e'e.meet(e) \wedge date(e, tomorrow) \wedge meet(e') \wedge date(e', monday) \wedge e = e' \tag{9.4}$$

On obtient bien cette représentation si la négociation est terminée et si le premier locuteur accepte cette proposition. Si tel n'est pas le cas et s'il refuse la négociation avec une réponse comme A_3, il s'agit cette fois de reprendre la représentation pour en enlever une partie. En repartant de la version non simplifiée (9.4), il suffit de supprimer la dernière égalité. Dans l'hypothèse où nous réalisons les simplifications au fur et à mesure, il convient de reconstruire la représentation (9.3). On voit bien ici que la nature du rattachement de la prise de parole au reste de la structure autorise ou non l'utilisation de processus spécifiques.

On retrouvera dans (Tiv 2016) une présentation plus élaborée de cette problématique, ainsi que plusieurs exemples. Il n'est pas question ici de soumettre une proposition de modélisation, mais bien de lier ces exemples avec les propositions précédentes de ce document. Il semble ici que traiter formellement des dialogues n'est pas traiter des discours et il est nécessaire d'intégrer à leur représentation leurs spécificités. Il ne faudrait pas comprendre que nous avançons un contre-argument à la modélisation du projet SLAM. Dans ce cadre, nous faisons une hypothèse importante quant au rôle du psychologue qui n'est pas un interlocuteur normal et qui a pour fonction, rappelons-le, de maintenir l'interaction avec le patient. Dans la quasi intégralité de ses prises de paroles, le psychologue va dans le sens du patient.

Pour passer d'une représentation de la sémantique à la Montague à une représentation du dialogue, nous avons besoin de prendre en compte le contexte général. Par ailleurs, nous devons disposer d'une représentation intermédiaire dans laquelle une négociation est possible avant de transférer le contenu sémantique à la représentation finale. Dans cet espace la cohérence n'est pas garantie et il est possible de revenir sur ce qui a été introduit. Nous nous rapprochons grandement des arguments en faveur de la dynamicité de la sémantique. Par ailleurs, le type de rattachement semble permettre des opérations de nature différente dans une version automatique et compositionnelle. Il est alors nécessaire de disposer d'un moyen d'identifier le type de rattachement au moins par les marqueurs explicites, mais très probablement en faisant appel aux marqueurs implicites (qui sont particulièrement difficiles à identifier).

Enfin, comme le dernier exemple nous le montre, il est également nécessaire d'avoir une modélisation du savoir partagé. Pour maintenir la cohérence générale de l'interaction, des inférences sont nécessaires : il faut par exemple pouvoir inférer d'une connaissance globale que « demain » n'est pas lundi. On voit également que disposer d'une formalisation des dialogues s'inscrirait naturellement dans la perspective du projet SLAM sur les exemples pathologiques comme cas particuliers des exemples généraux. Pour parvenir à mobiliser tous ces aspects, il nous faudrait disposer d'une grammaire sémantique à large couverture.

Ces deux derniers paragraphes reprennent la plupart des éléments dont nous avons discuté et plaident une nouvelle fois pour l'utilisation de TTDL qui a montré sa grande flexibilité tout en s'inscrivant dans la perspective de la sémantique montagovienne, ainsi qu'en utilisant des ACG qui ne dépendent d'aucune théorie et permettent de modéliser celle choisie pour la représentation.

9.5 Conclusion

Nous arrivons au terme de la présentation de nombre de nos travaux de recherche et de nos perspectives. Notre champ de recherche se situe en linguistique computationnelle. Il s'agit d'utiliser des outils théoriques et formels pour modéliser des phénomènes linguistiques.

Ce document s'est ouvert sur une longue contextualisation de ce type de travaux. L'occasion de donner une perspective épistémologique sur les questions qui traversent notre recherche est rare. Nous sommes revenus sur les notions de langue et langage pour arriver à définir ce que sont les objets sur lesquels nous travaillons. On constate que la logique est intrinsèquement présente dans la langue naturelle et que l'objectif est d'isoler ces éléments pour les utiliser dans d'autres tâches. En partant de ce postulat, nous avons donné une perspective historique sur l'émergence de ces questions et leurs évolutions dans le temps, jusqu'à aujourd'hui. Ces problématiques appartiennent selon nous simultanément à la linguistique, la logique et l'informatique. Il est toujours difficile de les positionner exclusivement sur l'un de ces domaines, mais ce qui rapproche d'une discipline plutôt que d'une autre est naturellement le prisme par lequel les questions sont abordées. Dans notre travail, nous nous situons très clairement du côté de l'informatique et de la logique, tout en ayant un profond intérêt pour les aspects linguistiques. De manière très simplifiée, nous pouvons situer ces travaux dans la continuité des propositions de Montague, en intégrant les nombreuses évolutions établies depuis cette époque.

Pendant notre thèse, nous avons d'abord travaillé sur le transfert de l'information structurelle de la modélisation syntaxique vers le niveau sémantique. Pour cela nous avons modélisé les relations de la théorie générative dans un cadre logique inspiré des grammaires catégorielles pour synchroniser un calcul sémantique. Le calcul utilise à la fois des connecteurs commutatifs et non-commutatifs. Nous avons montré la normalisation faible du calcul. Nous avons ensuite étendu le cadre en utilisant l'ensemble des connecteurs en leur attribuant des rôles différents dans l'analyse, tout en introduisant une interface syntaxe-sémantique simplifiée.

Dans la suite de nos travaux, nous nous sommes concentrés sur la modélisation sémantique, en particulier à partir du résultat proposé dans (de Groote 2006). Cette perspective permet de disposer d'un cadre dans lequel nous pouvons réconcilier les approches compositionnelles, inspirées de Frege, avec les modélisations linguistiques de la langue naturelle. Par exemple, il devient possible d'intégrer le traitement de phénomènes dynamiques dans le calcul originel de Montague, et d'aller jusqu'à la modélisation de phénomènes contextualisées complexes. Dans sa thèse que nous avons co-encadrée avec Philippe de Groote, Sai Qian a en particulier travaillé à la modélisation de plusieurs phénomènes comme l'accessibilité de variables sous phénomènes de double négation, ou encore la subordination modale. Ces phénomènes sont très présents dans la langue et interagissent profondément avec la modélisation.

Il apparaît rapidement que proposer des traitements pour ces phénomènes ne construit pas un fragment cohérent qui représente la langue. L'ensemble des traitements ont eux-mêmes des interactions. Aussi, il convient de définir un calcul dans lequel la composi-

tion des représentations est possible. C'est la proposition faite par Jirka Maršík dans sa thèse, (Maršík 2016). Il a abordé cette question en s'inspirant de propriétés des langages de programmation en reprenant les effets algébriques. Il a ainsi introduit un calcul qui simule les appels par nom et les appels par valeur. Ce calcul vérifie la confluence et la terminaison, ce qui assure l'existence d'une forme normale unique. Une certaine expertise reste nécessaire pour développer les grammaires, ce qui explique qu'il ne soit pas aisé de proposer rapidement une grammaire sémantique d'un large fragment du français ou de l'anglais.

Dans la seconde partie du manuscrit, nous avons abordé nos recherches relatives au discours. Ces éléments proviennent principalement du projet SLAM dont nous avons assuré la responsabilité scientifique. Le cœur du projet est constitué d'entretiens entre des patients schizophrènes et des psychologues. L'enjeu en est d'analyser les interactions linguistiques pour identifier des troubles du langage manifestes de troubles de la pensée.

Partant des premiers résultats, nous avons décidé de constituer une ressource plus large et d'y ajouter des annotations à différents niveaux. L'une des préoccupations était d'identifier les niveaux linguistiques relevant d'une défaillance qui pouvait s'inscrire dans la symptomatologie de la pathologie. Pour cela nous avons développé un outil qui automatise l'annotation (grâce à d'autres outils au niveau de l'état de l'art), ainsi qu'une partie importante de l'analyse quantitative. Nous avons ainsi pu montrer une apparition supérieure de disfluences chez les patients schizophrènes par rapport aux autres interlocuteurs. Par ailleurs, nous avons cherché à outiller méthodologiquement la définition et le déploiement de campagnes d'annotations manuelles sur notre ressource.

Nous avons pu mettre en évidence l'existence de phénomènes spécifiques à la pathologie que nous avons interprétés par la mobilisation de formalismes classiques de représentation du discours. Pour cela nous avons utilisé une adaptation à notre étude de la SDRT. Il apparaît que les schizophrènes tentent de construire des représentations de l'interaction dans lesquelles des références ne sont pas interprétables, car les représentations sont structurellement inconsistantes. Nous avons également explicité comment les interlocuteurs possèdent des représentations mentales de l'échange différentes. Ces éléments nous renseignent sur les processus mis en œuvre dans le dialogue, et ce au travers des relations linguistiques sémantico-pragmatiques.

Un dernier aspect concernant le discours est l'utilisation de sa représentation dans d'autres tâches du TAL, par exemple l'extraction d'informations. Si le projet SLAM donne un cadre d'application à la modélisation sémantique, il est intéressant de tenter de la mobiliser par exemple pour la fouille de données. Il s'agit d'utiliser des théories déjà établies et de leur donner comme entrées des éléments issus de la modélisation du discours. Nous avons ainsi travaillé à utiliser les marqueurs de relations discursives explicites. Le manque de ressources rend la tâche relativement compliquée, en particulier pour l'évaluation.

Ce dernier point clôt la présentation des résultats et permet d'ouvrir vers la description des perspectives de recherche. Le projet se situe très clairement pour une partie du côté de la production automatique de représentations sémantiques, et pour l'autre part sur l'application des modélisations sur des données réelles. Nous revenons sur les grandes

perspectives pour la suite de nos recherches.

Le premier aspect qu'il est nécessaire de poursuivre est le développement de l'outil `SLAMtk`. Il correspond aux besoins du projet SLAM actuel. Il est en plus possible de capitaliser en étendant ses possibilités de traitements et d'en tirer profit dans des contextes différents. Nous avons ainsi pu proposer d'ajouter un certain nombre d'analyses quantitatives supplémentaires. Par ailleurs, il faut traiter plus précisément de la syntaxe, ainsi qu'avoir recours à des résultats récents pour la transcription automatique.

Il semble également important de poursuivre sur la modélisation formelle. Les premiers résultats sont intéressants, mais il faut d'une part les confronter à de nouveaux corpus qui recouvrent notamment d'autres pathologies, nous prévoyons pour cela de poursuivre la constitution de la ressource, et d'autre part de travailler plus précisément aux enjeux de la modélisation. Nous devons en particulier étudier à nouveau la notion de processus cognitifs et comment nous les intégrons au calcul. Il apparaît alors que nous pouvons en isoler certains et leur donner un rôle dans le passage des troubles de la pensée aux troubles du langage.

Comme évoqué précédemment, l'enjeu est de faire le lien entre la définition d'outils théoriques et formels pour la sémantique des langues naturelles et leur utilisation dans des contextes spécifiques. La perspective principale est donc le développement d'une grammaire sémantique pour le français à large couverture. C'est le point structurant pour la suite de nos recherches. L'élément principal est de proposer un lexique à large couverture pour le français (et pour l'anglais). Plusieurs pistes pour développer des premières versions ont d'ores et déjà été mises en avant, mais le problème demeure ardu pour parvenir à couvrir correctement tous les aspects.

Enfin, même si disposer d'une grammaire sémantique pour le français permettrait de faire un pas important sur le lien entre représentation sémantique logique et SLAM, il convient également de définir un cadre logique formel pour rendre compte des dialogues. Ainsi la chaîne de traitements serait complète pour passer de l'un à l'autre. Des premiers éléments sont en cours de développement, mais une théorisation fine nécessite de considérer de nombreux cas.

Ces quatre derniers paragraphes dressent les perspectives de notre recherche pour les prochaines années. Il s'agira alors de parvenir à réconcilier la vision calculatoire de la modélisation sémantique avec ses applications dans des perspectives des sciences cognitives.

Bibliographie

Abdelaziz, Ahmed H. et al. (2015). "Uncertainty propagation through deep neural networks". In : *Interspeech 2015*. Dresden, Germany. HAL archive ouverte : hal-01162550 (cf. p. 117).

Abeillé, A., L. Clément et F. Toussenel (2003). "Building a treebank for French". In : *Treebanks*. Sous la dir. d'A. Abeillé. Kluwer, Dordrecht (cf. p. 25).

Abeillé, Anne et Benoit Crabbé (2013). "Vers un treebank du français parlé". In : *Traitement Automatique des Langues Naturelles (TALN)*. Les Sables d'Olonne, France, p. 174–187 (cf. p. 116).

Afantenos, Stergos et al. (2012). "An empirical resource for discovering cognitive principles of discourse organisation : the ANNODIS corpus (regular paper)". In : *Language Resources and Evaluation Conference (LREC), Istanbul, Turkey, 23/05/2012-25/05/2012*. Sous la dir. de Nicoletta Calzolari. http ://www.elra.info : European Language Resources Association (ELRA), (on line) (cf. p. 125).

Ajdukiewicz, Kazimierz (1967). "Logika pragmatyczna". In : (cf. p. 16).

Allen, James F (1983). "Maintaining knowledge about temporal intervals". In : *Communications of the ACM* 26.11, p. 832–843 (cf. p. 24).

— (1984). "Towards a general theory of action and time". In : *Artificial intelligence* 23.2, p. 123–154 (cf. p. 24).

Amblard, Maxime (2007). "Calculs de représentations sémantiques et syntaxe générative : les grammaires minimalistes catégorielles". Thèse de doct. Université Sciences et Technologies - Bordeaux I. HAL archive ouverte : tel-00185844 (cf. p. 8, 22, 29, 31, 32, 33, 35, 37, 43).

— (2011a). "Encoding Phases using Commutativity and Non-commutativity in a Logical Framework". In : *Logical Aspect of Computational Linguistic*. Sous la dir. de Sylvain Pogodalla et Jean-Philippe Prost. T. 6736. Lecture Notes in Artificial Intelligence. ISBN : 978-3-642-22220-7 The original publication is available at www.springerlink.com. Montpellier, France : Springer, p. 1–16. DOI : 10.1007/978-3-642-22221-4_1. HAL archive ouverte : hal-00601621 (cf. p. 29).

— (2011b). "Minimalist Grammars and Minimalist Categorial Grammars, definitions toward inclusion of generated languages". In : *Logic and Grammar*. Sous la dir. de Sylvain Pogodalla, Myriam Quatrini et Christian Retoré. T. 6700. LNCS/LNAI. springer, p. 61–80. DOI : 10.1007/978-3-642-21490-5_4. HAL archive ouverte : hal-00617040 (cf. p. 35, 43).

— (2013). "Real Humans : des machines qui parlent comme des hommes, ou presque..." In : *Interstices*. HAL archive ouverte : hal-01281272 (cf. p. 1).

Amblard, Maxime (2015). “La non-commutativité comme argument linguistique : modéliser la notion de phase dans un cadre logique”. In : *Traitement Automatique des Langues* 56.1, p. 91–115. HAL archive ouverte : `hal-01188669` (cf. p. 29, 32, 33, 41).

— (en soumission). “Pour un TAL éthique”. In : *Traitement Automatique des Langues* 57.2 (cf. p. 7).

Amblard, Maxime et Karën Fort (2014). “Étude quantitative des disfluences dans le discours de schizophrènes : automatiser pour limiter les biais”. In : *TALN - Traitement Automatique des Langues Naturelles*. Marseille, France, p. 292–303. HAL archive ouverte : `hal-01054391` (cf. p. 97).

Amblard, Maxime, Karën Fort, Caroline Demily et al. (2015). “Analyse lexicale outillée de la parole transcrite de patients schizophrènes”. In : *Traitement Automatique des Langues*. Natural Language Processing and Cognition 55.3, p. 91–115. HAL archive ouverte : `hal-01188677` (cf. p. 84, 97, 102).

Amblard, Maxime, Karën Fort, Michel Musiol et al. (2014). “L’impossibilité de l’anonymat dans le cadre de l’analyse du discours”. In : *Journée ATALA éthique et TAL*. Paris, France. HAL archive ouverte : `hal-01079308` (cf. p. 7, 97, 100, 118).

Amblard, Maxime, Alain Lecomte et Christian Retoré (2010). “Categorial Minimalist Grammar : From Generative Syntax To Logical Form”. In : *Linguistic Analysis* 36.1–4, p. 273–306. HAL archive ouverte : `hal-00545748` (cf. p. 37).

Amblard, Maxime, Michel Musiol et Manuel Rebuschi (2011). “Une analyse basée sur la S-DRT pour la modélisation de dialogues pathologiques”. In : *Traitement Automatique des Langues Naturelles - TALN 2011*. Sous la dir. de Mathieu Lafourcade et Violaine Prince. Montpellier, France : Laboratoire d’Informatique de Robotique et de Microélectronique, p. 6. HAL archive ouverte : `hal-00601622` (cf. p. 89).

— (2012). “Schizophrénie et Langage : Analyse et modélisation. De l’utilisation des modèles formels en pragmatique pour la modélisation de discours pathologiques”. In : *Congrès MSH 2012*. Caen, France. HAL archive ouverte : `hal-00761540` (cf. p. 89).

— (2014). “L’interaction conversationnelle à l’épreuve du handicap schizophrénique.” In : *Recherches sur la philosophie et le langage* 31, p. 1–21. HAL archive ouverte : `hal-00955660` (cf. p. 91).

Amblard, Maxime et Sylvain Pogodalla (2014). “Modeling the Dynamic Effects of Discourse : Principles and Frameworks”. In : *Dialogue, Rationality, and Formalism*. Sous la dir. de Manuel Rebuschi et al. T. 3. Interdisciplinary Works in Logic, Epistemology, Psychology and Linguistics. Logic, Argumentation & Reasoning. Springer, p. 247–282. DOI : `10.1007/978-3-319-03044-9\_12`. HAL archive ouverte : `hal-00737765` (cf. p. 24, 45, 46, 88).

Amblard, Maxime et Christian Retoré (2014). “Partially Commutative Linear Logic and Lambek Caculus with Product : Natural Deduction, Normalisation, Subformula Property”. In : *IfColog Journal of Logics and their Applications (FLAP)* 1.1, p. 53–94. HAL archive ouverte : `hal-01071642` (cf. p. 35, 43).

Aronson, A. (2014). *MetaMap - A Tool For Recognizing UMLS Concepts in Text*. http ://metamap.nlm.nih.gov (cf. p. 111).

Asher, Nicholas (1993). *Reference to Abstract Objects in Discourse : A Philosophical Semantics for Natural Language Metaphysics*. T. 50. SLAP. Kluwer (cf. p. 2).

Asher, Nicholas et Alex Lascarides (1998). "The Semantics and Pragmatics of Presupposition". In : *Journal of Semantics* 15.2, p. 239–299 (cf. p. 46).

— (2003). *Logics of Conversation*. Studies in Natural Language Processing. Cambridge University Press. ISBN : 9780521650588 (cf. p. 3, 6, 26, 52, 87, 88, 89, 110).

Asher, Nicholas et Sylvain Pogodalla (2010). "A Montagovian Treatment of Modal Subordination". In : *Proceedings of Semantics and Linguistic Theory (SALT) 20*. Sous la dir. de Nan Li et David Lutz. Linguistic Society of America. Vancouver : *e*Language, p. 387–405 (cf. p. 55).

Bach, Emmon (1986). "Natural Language Metaphysics". In : *Studies in Logic and the Foundations of Mathematics* 114, p. 573–595 (cf. p. 2).

— (1989). *Informal lectures on formal semantics*. Suny Press (cf. p. 23).

Baldridge, Jason et Alex Lascarides (2005). "Probabilistic head-driven parsing for discourse structure". In : *Proceedings of the Ninth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, p. 96–103 (cf. p. 111).

Bar-Hillel, Yehoshua (1953). "A quasi-arithmetical notation for syntactic description". In : *Language* 29.1, p. 47–58 (cf. p. 16).

Barker, Chris (2002). "Continuations and the nature of quantification". In : *Natural language semantics* 10.3, p. 211–242. DOI : 10.1023/A:1022183511876 (cf. p. 46).

— (2004). "Continuations in natural language". In : *CW* 4, p. 1–11 (cf. p. 46).

Barras, Claude et al. (1998). "Transcriber : a Free Tool for Segmenting, Labeling and Transcribing Speech". In : *International Conference on Language Resources and Evaluation (LREC)*. Granada, p. 1373–1376 (cf. p. 79).

Barsalou, L.W. (1992). "Frames, concepts, and conceptual fields". In : *Frames, Fields and Contrasts. New Essays in Semantic and Lexical Organization*. S. 21-74. (cf. p. 111, 126).

Barwise, Jon (1987). *Noun phrases, generalized quantifiers and anaphora*. Springer (cf. p. 26).

Beaver, David Ian (1997). "Presupposition". In : *Handbook of Logic and Language*. Sous la dir. de Johan van Benthem et Alice ter Meulen. Elsevier Science Publishers. Chap. 17, p. 939–1008 (cf. p. 124).

— (2002). "Presupposition Projection in DRT : A critical assessment". In : *The construction of meaning*. Citeseer (cf. p. 25).

Benthem, Johan van (2013). *The logic of time : a model-theoretic investigation into the varieties of temporal ontology and temporal discourse*. T. 156. Springer Science & Business Media (cf. p. 24).

Benzitoun, Christophe, Karën Fort et Benoît Sagot (2012). "TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe". In : *Traitement Automatique des Langues Naturelles (TALN)*. Grenoble, France, p. 99–112. Anthologie ACL : F12-2008 (cf. p. 99).

Birkhoff, Garrett et John Von Neumann (1936). "The Logic of Quantum Mechanics". In : *Annals of Mathematics* 37.4, p. 823–843 (cf. p. 17).

Blanche-Benveniste, Claire (1997). *Approches de la langue parlée en français*. Collection L'Essentiel français. Gap, France : Ophrys (cf. p. 79).

Blanqui, Frédéric (2000). "Termination and confluence of higher-order rewrite systems". In : *Rewriting Techniques and Applications*. Springer, p. 47–61 (cf. p. 66).

Blanqui, Frédéric, Jean-Pierre Jouannaud et Mitsuhiro Okada (2002). "Inductive-data-type systems". In : *Theoretical Computer Science* 272.1, p. 41–68 (cf. p. 66).

Bloom, Floyd E (1993). "Advancing a neurodevelopmental origin for schizophrenia". In : *Archives of general psychiatry* 50.3, p. 224 (cf. p. 90).

Bos, Johan (2008). "Wide-Coverage Semantic Analysis with Boxer". In : *Semantics in Text Processing. STEP 2008 Conference Proceedings*. Sous la dir. de Johan Bos et Rodolfo Delmonte. Research in Computational Semantics. College Publications, p. 277–286 (cf. p. 125).

Bos, Johan et al. (2016). "Handbook of Linguistic Annotation". In : sous la dir. de Nancy Ide et James Pustejovsky. Berlin : Springer. Chap. The Groningen Meaning Bank (cf. p. 122, 125).

Boula de Mareüil, Philippe et al. (2013). "Une étude quantitative des marqueurs discursifs, disfluences et chevauchements de parole dans des interviews politiques". In : *TIPA. Travaux interdisciplinaires sur la parole et le langage [En ligne]* 29 (cf. p. 100).

Bowie, Christopher R. et Philip D. Harvey (2006). "Administration and interpretation of the Trail Making Test". In : *Nature Protocols* 1.5, p. 2277–2281. ISSN : 1754-2189. DOI : 10.1038/nprot.2006.390 (cf. p. 82).

Carlson, Lynn et al. (2002). *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania (cf. p. 111).

Carnap, Rudolf (1947). *Meaning and Necessity* (cf. p. 15).

— (1952). "Meaning postulates". In : *Philosophical studies* 3.5, p. 65–73 (cf. p. 15).

Cerisara, Christohe, Odile Mella et Dominique Fohr (2009). "JTrans, an open-source software for semi-automatic text-to-speech alignment". In : *Proc. of INTERSPEECH*. Brighton, UK (cf. p. 117).

Chaika, Elaine (1974). "A linguist looks at "schizophrenic" language". In : *Brain and Language* 1.3, p. 257–276 (cf. p. 87).

Charlow, Simon (2014). "On the semantics of exceptional scope". Thèse de doct. New York University (cf. p. 63).

Chomsky, N. et M.P. Schützenberger (1963). "The Algebraic Theory of Context-Free Languages". In : *Computer Programming and Formal Systems*. Sous la dir. de P. Braffort et D. Hirschberg. T. 35. Studies in Logic and the Foundations of Mathematics. Elsevier, p. 118–161. DOI : 10.1016/S0049-237X(08)72023-8 (cf. p. 22).

Chomsky, Noam (1957). *Syntactic Structures*. Mouton & Co. (cf. p. 2, 5, 11, 22, 23, 30).

— (1981). "Lectures on Government and Binding". In : *Dordrecht : Foris* (cf. p. 2, 5, 22).

— (1993). *Lectures on government and binding : The Pisa lectures*. T. 9. Walter de Gruyter (cf. p. 22).

— (1995). *The minimalist program*. T. 28. Cambridge Univ Press (cf. p. 2, 5, 30).

— (1999). *Derivation by phase*. ms, MIT (cf. p. 5, 22, 29, 37).

Chomsky, Noam et Robert DiNozzi (1972). *Language and mind*. Harcourt Brace Jovanovich New York (cf. p. 22).

Church, Alonzo (1940). "A formulation of the simple theory of types". In : *The journal of symbolic logic* 5.2, p. 56–68 (cf. p. 18, 47).

— (1951). "A formulation of the logic of sense and denotation". In : *Structure, Method, and Meaning : Essays in Honor of Henry M. Sheffer*, p. 3–24 (cf. p. 18).

Clark, Herbert et Eve Clark (1977). *Psychology and language : An introduction to psycholinguistics*. Harcourt Brace Jovanovich (New York) (cf. p. 120).

Clark, Herbert, Robert Schreuder et Samuel Buttrick (1983). "Common ground at the understanding of demonstrative reference". In : *Journal of Verbal Learning and Verbal Behavior* 22.2, p. 245–258. ISSN : 0022-5371. DOI : 10.1016/S0022-5371(83)90189-5 (cf. p. 120).

Clark, Stephen (2015). "Vector Space Models of Lexical Meaning". In : sous la dir. de Shalom Lappin et Chris Fox. Wiley-Blackwell. Chap. 16, p. 493–522 (cf. p. 122).

Committee, Automatic Language Processing Advisory (1966). *Language and Machines*. National Academy of Sciences et National Research Council (cf. p. 18).

Constant, Matthieu et Anne Dister (2010). "Automatic detection of disfluencies in speech transcriptions". In : *Spoken Communication*. Sous la dir. de M. Pettorino et al. T. 1. Cambridge Scholars Publishing, p. 259–272. HAL archive ouverte : hal-00636983 (cf. p. 97).

Crabbé, Marcel (2000). *Une introduction à la logique du premier ordre : lois logiques, raisonnements valides, paradoxes, notion de modèle, calcul des séquents, théorème de complétude*. textes en ligne (cf. p. 1, 13).

Curry, Haskell (1934). "Functionality in Combinatory Logic." In : *Proceedings of the National Academy of Sciences of the United States of America*. T. 20. 11, p. 584–590 (cf. p. 2).

Danlos, Laurence et al. (2012). "Vers le FDTB : French Discourse Tree Bank". In : *JAD'12 - Journée Atala Discours*. ATALA et revue Discours. Paris, France. HAL archive ouverte : hal-00704705 (cf. p. 125).

Davidson, Donald (1965). "Theories of Meaning and Learnable Languages". In : *Proceedings of the International Congress for Logic, Methodology, and Philosophy of Science*. North-Holland, p. 3–17 (cf. p. 51).

— (1967). "Truth and meaning". In : *Synthese* 17.1, p. 304–323 (cf. p. 41, 51).

— (2001). *Essays on actions and events*. T. 1. Oxford University Press (cf. p. 45).

de Groote, Philippe (1996). "Partially commutative linear logic : sequent calculus and phase semantics". In : *Third Roma Workshop : Proofs and Linguistics Categories – Applications of Logic to the analysis and implementation of Natural Language*. Sous la dir. de Vito Michele Abrusci et Claudia Casadio. Bologna :CLUEB, p. 199–208 (cf. p. 29, 34).

— (2001). "Towards abstract categorial grammars". In : *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p. 252–259 (cf. p. 5, 45, 123).

de Groote, Philippe (2006). "Towards a Montagovian account of dynamics". In : *Proceedings of Semantics and Linguistic Theory (SALT) 16*. Sous la dir. de Masayuki Gibson et Jonathan Howell (cf. p. v, vii, 3, 5, 8, 26, 40, 43, 46, 47, 48, 50, 121, 129).

— (2010). *Dynamic Semantics and Discourse*. Lectures at the International Conference on Semantics and Formal Modelling, JSM'10 (cf. p. 49).

de Groote, Philippe et Makoto Kanazawa (2013). "A Note on Intensionalization". In : *Journal of Logic, Language and Information* 22.2, p. 173–194 (cf. p. 63).

de Groote, Philippe et Ekaterina Lebedeva (2010). "Presupposition accommodation as exception handling". In : *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, p. 71–74 (cf. p. 50).

De La Clergerie, Éric et al. (2009). "FRMG : évolutions d'un analyseur syntaxique TAG du français". In : *Journée de l'ATALA sur : Quels analyseurs syntaxiques pour le français ?* Sous la dir. d'Éric Villemonte de la Clergerie et Patrick Paroubek. Journée de l'ATALA organisée conjointement à la conférence IWPT 2009. ATALA. Paris, France. HAL archive ouverte : `inria-00553260` (cf. p. 116).

de Saussure, Ferdinand (1916). *Cours de linguistique générale*. Otto Harrassowitz Verlag (cf. p. 1, 13).

Dekker, Paul (1994). "Predicate Logic with Anaphora". In : *Proceedings of the Fourth Semantics and Linguistic Theory Conference (SALT)*. Sous la dir. de Lynn Santelmann et Mandy Harvey. DMLL Publications, Cornell University (cf. p. 3).

Denis, Pascal et Benoît Sagot (2009). "Coupling an Annotated Corpus and a Morphosyntactic Lexicon for State-of-the-Art POS Tagging with Less Human Effort". In : *Pacific Asia Conference on Language Information and Computing (PACLIC)*. Hong-Kong (cf. p. 99).

Deulofeu, José et al. (2010). "Depends on What the French Say Spoken Corpus Annotation with and Beyond Syntactic Functions". In : *Proceedings of the Fourth Linguistic Annotation Workshop*. LAW IV '10. Uppsala, Sweden : Association for Computational Linguistics, p. 274–281. ISBN : 978-1-932432-72-5 (cf. p. 116).

Dowty, David (1979). *Word meaning and Montague grammar : The semantics of verbs and times in generative semantics and in Montague's PTQ*. T. 7. Springer (cf. p. 124).

DSMIV (1994). *Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR*. Fourth Edition (Text Revision). T. 30. Washington DC : American Psychiatric Association (cf. p. 87).

Dumais, Susan (2003). "Data-driven approaches to information access". In : *Cognitive Science* 27, p. 491–524 (cf. p. 122).

Dumais, Susan T. (2004). "Latent semantic analysis". In : *Annual Review of Information Science and Technology* 38.1, p. 188–230 (cf. p. 122).

Eijck, Jan van (1999). "Axiomatising Dynamic Logics for Anaphora". In : *Journal of Language and Computation*. T. 1 (cf. p. 26).

Erk, Katrin (2012). "Vector Space Models of Word Meaning and Phrase Meaning : A Survey." In : *Language and Linguistics Compass* 6.10, p. 635–653 (cf. p. 122).

Eshkol-Taravella, Iris et al. (2014). *Procédure d'anonymisation et traitement automatique : l'expérience d'ESLO*. Journée d'études ATALA, Ethique et TAL. Paris (cf. p. 96).

Farkas, Donka F. (1981). "Quantifier Scope and Syntactic Islands". In : *Papers from the Seventeenth Regional Meeting of the Chicago Linguistic Society (CLS 17)*. Sous la dir. de Roberta Hendrick, Carrie Masek et Mary Frances Miller, p. 59–66 (cf. p. 70).

Feferman, Anita Burdman et Solomon Feferman (2004). *Alfred Tarski : life and logic*. Cambridge University Press (cf. p. 20).

Fort, Karën (2012). "Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus". Thèse de doct. Université Paris XIII, LIPN, INIST-CNRS. HAL archive ouverte : tel-00797760 (cf. p. 103, 106, 107).

Fort, Karën et Alain Couillault (2016). "Yes, We Care ! Results of the Ethics and Natural Language Processing Surveys". In : *Language Resources and Evaluation Conference (LREC)*. Portorož, Slovénie (cf. p. 7).

Frege, Gottlob (1879). "Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens. Halle a. S. : Louis Nebert. Translation : Concept Script, a formal language of pure thought modelled upon that of arithmetic, by S. Bauer-Mengelberg". In : *VAN HEIJENOORT. From Frege to Gödel : a Source Book in Mathematical Logic* 1931 (cf. p. 14).

— (1892). "Ausführungen über Sinn und Bedeutung". In : *Nachgelassene Schriften*, p. 128–135 (cf. p. 2, 14).

— (1994). "Über sinn und bedeutung". In : *Wittgenstein Studien* 1.1 (cf. p. 14).

Fromkin, Victoria A. (1975). "A linguist looks at "a linguist looks at 'schizophrenic language"'". In : *Brain and Language* 2, p. 498–503. ISSN : 0093-934X. DOI : 10.1016/S0093-934X(75)80087-3 (cf. p. 87).

Ganter, B., C. Franzke et R. Wille (2012). *Formal Concept Analysis : Mathematical Foundations*. Springer Berlin Heidelberg. ISBN : 9783642598302 (cf. p. 7).

Ganter, B. et S.O. Kuznetsov (2001). "Pattern Structures and Their Projections". In : *Conceptual Structures : Broadening the Base, Proceedings of the 9th International Conference on Conceptual Structures, ICCS 2001, Stanford, CA*. Sous la dir. de H.S. Delugach et G. Stumme. Lecture Notes in Computer Science 2120. Springer, p. 129–142 (cf. p. 109, 112).

Ganter, B. et R. Wille (1999). *Formal Concept Analysis, Mathematical Foundations*. Springer (cf. p. 109, 111).

Gardent, Claire (2008). "Integrating a Unification-based Semantics in a Large Scale Lexicalised Tree Adjoining Grammar for French". In : *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*. COLING '08. Manchester, United Kingdom : Association for Computational Linguistics, p. 249–256. ISBN : 978-1-905593-44-6 (cf. p. 123, 124).

Gardent, Claire et Yannick Parmentier (2005). "Large Scale Semantic Construction for Tree Adjoining Grammars". In : *Proceedings of the 5th International Conference on Logical Aspects of Computational Linguistics*. LACL'05. Bordeaux, France : Springer-

Verlag, p. 131–146. ISBN : 3-540-25783-7, 978-3-540-25783-7. DOI : 10.1007/11422532_9 (cf. p. 123).

Geurts, Bart (1999). *Presuppositions and Pronouns*. Current Research in the Semantics/Pragmatics Interface. Elsevier (cf. p. 25).

— (2002). "Donkey business". In : *Linguistics and philosophy* 25.2, p. 129–156 (cf. p. 24, 25).

Gödel, Kurt (1931). "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I". In : *Monatshefte für mathematik und physik* 38.1, p. 173–198 (cf. p. 16).

Groenendijk, Jeroen et Martin Stokhof (1990). "Dynamic Montague grammar". In : *Proceedings of the Second Symposion on Logic and Language*, p. 3–48 (cf. p. 25, 26).

— (1991). "Dynamic predicate logic". In : *Linguistics and philosophy* 14.1, p. 39–100 (cf. p. 3, 25, 111).

Grosz, Barbara J. et Candace L. Sidner (1986). "Attention, Intentions, and the Structure of Discourse". In : *Computational Linguistics* 12.3, p. 175–204 (cf. p. 3).

Grouin, Cyril et Pierre Zweigenbaum (2013). "Automatic De-Identification of French Clinical Records : Comparison of Rule-Based and Machine-Learning Approches". In : *Stud Health Technol Inform, Proc of MEDINFO*. T. 192. Copenhagen, Denmark, p. 476–80. DOI : 10.3233/978-1-61499-289-9-476 (cf. p. 95).

Hacquard, Valentine (2006). "Aspects of modality". Thèse de doct. Massachusetts Institute of Technology (cf. p. 54).

Hamana, Makoto (2007). "Higher-order semantic labelling for inductive datatype systems". In : *Proceedings of the 9th ACM SIGPLAN international conference on Principles and practice of declarative programming*. ACM, p. 97–108 (cf. p. 66).

Harris, Z. (1954). "Distributional structure". In : *Word* 10.23, p. 146–162 (cf. p. 11, 118, 122).

Heiden, Serge (2010). "The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme". In : *24th Pacific Asia Conference on Language, Information and Computation*. Sous la dir. de Ryo Otoguro et al. Sendai, Japan : Institute for Digital Enhancement of Cognitive Development, Waseda University, p. 389–398. HAL archive ouverte : halshs-00549764 (cf. p. 118).

Heiden, Serge, Jean-Philippe Magué et Bénédicte Pincemin (2010). "TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement". In : *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*. Sous la dir. de Luca Giuliano Sergio Bolasco Isabella Chiari. T. 2-3. Rome, Italie : Edizioni Universitarie di Lettere Economia Diritto, p. 1021–1032. HAL archive ouverte : halshs-00549779 (cf. p. 99).

Heim, Irene (1982). "The Semantics of Definite and Indefinite Noun Phrases". Thèse de doct. University of Massachussetts, Amherst (cf. p. 25).

— (1983a). "File Change Semantics and the Familiarity Theory of Definiteness". In : *Meaning, Use and the Interpretation of Language*. Sous la dir. de Rainer Bäuerle, Christoph Schwarze et Arnim von Stechows. Reprinted in (Portner et Barbara H. Partee 2002). Walter de Gruyter & Co, p. 164–190 (cf. p. 25).

— (1983b). “File change semantics and the familiarity theory of definiteness”. In : *Formal Semantics*, p. 223–248 (cf. p. 25).
— (1990). “E-type pronouns and donkey anaphora”. In : *Linguistics and philosophy* 13.2, p. 137–177 (cf. p. 25).
Heim, Irene et Angelika Kratzer (1998). *Semantics in Generative Grammar*. Blackwell (cf. p. 30).
Hilbert, David (1902). “Sur les problèmes futurs des mathématiques”. In : *Compte rendu du deuxieme congres international des mathématiciens*, p. 58–114 (cf. p. 16).
— (1927). “The foundations of mathematics”. In : (cf. p. 16).
Hintikka, Jaakko (1957). “Modality as referential multiplicity”. In : *Eripainos Ajatus*. WSOY (cf. p. 54).
Hotho, Andreas, Andreas Nürnberger et Gerhard Paaß (2005). “A Brief Survey of Text Mining.” In : *Journal of Language Technologies and Computational Linguistics* 20.1, p. 19–62 (cf. p. 109).
Howard, William A. (1980). “The formulas-as-types notion of construction”. In : *To H. B. Curry : Essays on Combinatory Logic, Lambda Calculus, and Formalism*. Sous la dir. de J. P. Seldin et J. R. Hindley. Reprint of 1969 article. Academic Press, p. 479–490 (cf. p. 2, 25, 29).
Hyland, Martin, Gordon Plotkin et John Power (2006). “Combining effects : Sum and tensor”. In : *Theoretical Computer Science* 357.1, p. 70–99 (cf. p. 61).
Jackendoff, Ray (1972). *Semantic interpretation in generative grammar*. MIT press Cambridge, MA (cf. p. 30).
Jones, Simon L Peyton (2003). *Haskell 98 language and libraries : the revised report*. Cambridge University Press (cf. p. 63).
Kamp, Hans (1981). “A theory of truth and semantic representation”. In : *Formal Semantics*, p. 189–222 (cf. p. 25).
Kamp, Hans, Josef Genabith et Uwe Reyle (2011). “Discourse representation theory”. In : *Handbook of philosophical logic*, p. 125–394 (cf. p. 25).
Kamp, Hans et Uwe Reyle (1993). *From discourse to logic : introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory. Part 1*. From Discourse to Logic : Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. Kluwer Academic. ISBN : 9780792310273 (cf. p. 3, 46, 50, 87, 122).
Kleene, Stephen Cole (1943). “Recursive predicates and quantifiers”. In : *Transactions of the American Mathematical Society* 53.1, p. 41–73 (cf. p. 18).
Klop, Jan Willem, Vincent Van Oostrom et Femke Van Raamsdonk (1993). “Combinatory reduction systems : introduction and survey”. In : *Theoretical computer science* 121.1, p. 279–308 (cf. p. 66).
Kobele, Gregory (2006). “Generating Copies : An Investigation into Structural Identity in Language and Grammar”. Thèse de doct. University of California, Los Angeles (cf. p. 30).
Kratzer, Angelika (1977). “What ‘must’ and ‘can’ must and can mean”. In : *Linguistics and Philosophy* 1.3, p. 337–355 (cf. p. 55).

Kratzer, Angelika (1981). "The notional category of modality". In : *Words, worlds, and contexts*, p. 38–74 (cf. p. 55).

— (1986). "Conditionals". In : *Chicago Linguistics Society*. T. 22, p. 1–15 (cf. p. 55).

— (1991). "Modality". In : *Semantics : An international handbook of contemporary research*, p. 639–650 (cf. p. 55).

Lafon, Pierre (1980). "Sur la variabilité de la fréquence des formes dans un corpus". In : *Mots : Saussure, Zipf, Lagado, des méthodes, des calculs, des doutes et le vocabulaire de quelques textes politiques* 1.1, p. 127–165 (cf. p. 99).

Lambek, Joachim (1958). "The mathematics of sentence structure". In : *American mathematical monthly*, p. 154–170 (cf. p. v, vii, 5, 29).

— (1961). "On the Calculus of Syntactic Types". In : *Structure of Language and its Mathematical Aspects*. Sous la dir. de R. Jacobsen. Proceedings of Symposia in Applied Mathematics, XII. American Mathematical Society (cf. p. 29).

Landauer, Thomas K et Susan T. Dutnais (1997). "A solution to Plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge". In : *Psychological review* 104.2, p. 211–240 (cf. p. 122).

Lascarides, Alex et Nicholas Asher (2008). "Agreement and Disputes in Dialogue". In : *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics. Columbus, Ohio : Association for Computational Linguistics, 29–36. Anthologie ACL : `W08-0104` (cf. p. 125).

Lebedeva, Ekaterina (2012). "Expression de la dynamique du discours à l'aide de continuations". Thèse de doct. Université de Lorraine (cf. p. 49, 50, 63).

Lecomte, Alain (2005). "Categorial Grammar for Minimalism". In : *Language and Grammar : Studies in Mathematical Linguistics and Natural Language* CSLI Lecture Notes.168, p. 163–188 (cf. p. 33).

Lecomte, Alain et Christian Retoré (2001). "Extending Lambek grammars : a logical account of minimalist grammars". In : *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, ACL 2001*. Toulouse : ACL, p. 354–361 (cf. p. 33).

Leibniz, Gottfried Wilhelm (1685). *The Art of Dicovery*. T. 51. Wiener (cf. p. 14).

Levy Deborah L. abd Sereno, Anne B., Diane C. Gooding et Gilllian A. O'Driscolln (2010). "Eye Tracking Dysfunction in Schizophrenia : Characterization and Pathophysiology". In : *Curr Top Behav Neurosci.* 4, p. 311–347 (cf. p. 82).

Lindsey, DT et al. (1978). "Smooth-pursuit eye movements : a comparison of two measurement techniques for studying schizophrenia." In : *J Abnorm Psychol* 87.5, p. 481–96 (cf. p. 82).

Mann, William et Sandra Thompson (1988). "Rhetorical structure theory : Toward a functional theory of text organization". In : *Text* 8.3. Sous la dir. de LEditor Polanyi, p. 243–281. URL : `10.1515/text.1.1988.8.3.243` (cf. p. 3, 111).

Marcu, Daniel et Abdessamad Echihabi (2002). "An Unsupervised Approach to Recognizing Discourse Relations". In : *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. Association for Computational Linguistics, p. 368–375. DOI : `10.3115/1073083.1073145` (cf. p. 111).

Maršík, Jirka (2013). "Towards a Wide-Coverage Grammar : Graphical Abstract Categorial Grammars". Mém.de mast. Université de Lorraine (cf. p. 62).

— (2016). "Effects and Handlers in Natural Language". Thèse de doct. Université de Lorraine (cf. p. 6, 63, 66, 68, 72, 130).

Maršík, Jirka et Maxime Amblard (2013). "Integration of Multiple Constraints in ACG". In : *Logic and Engineering of Natural Language Semantics 10*. Kanagawa, Japan, p. 1–14. HAL archive ouverte : `hal-00869748` (cf. p. 61, 63).

— (2015). "Pragmatic Side Effects". In : *Redrawing Pragmasemantic Borders*. Groningen, Netherlands. HAL archive ouverte : `hal-01164729` (cf. p. 63).

— (2016). "Introducing a Calculus of Effects and Handlers for Natural Language Semantics". In : *Formal Grammar (in submission)* (cf. p. 6, 68).

Martin, Scott et Carl Pollard (2010). "Hyperintensional Dynamic Semantics : Analyzing Definiteness with Enriched Contexts". In : *Proceedings of Formal Grammar (FG) 15*. To appear in Lecture Notes in Computer Science (cf. p. 50).

— (2012). *A Higher-Order Theory ofPresupposition*. To appear in *Studia Logica* (cf. p. 50).

Mathet, Yann et Antoine Widlöcher (2011). "Stratégie d'exploration de corpus multi-annotés avec GlozzQL". In : *Actes de la 18e Conférence Traitement Automatique des Langues Naturelles (TALN'11), volume 2, papiers courts*. Sous la dir. de Mathieu Lafourcade et Violaine Prince. Montpellier, France, p. 143–148. HAL archive ouverte : `hal-01021846` (cf. p. 104).

Michaelis, Jens (2001). "Derivational Minimalism Is Mildly Context–Sensitive". In : *Logical Aspects of Computational Linguistics*. Sous la dir. de Michael Moortgat. T. 2014. Lecture Notes in Computer Science. Springer Berlin Heidelberg, p. 179–198. ISBN : 978-3-540-42251-8. DOI : `10.1007/3-540-45738-0_11` (cf. p. 35).

Moggi, Eugenio (1991). "Notions of computation and monads". In : *Information and computation* 93.1, p. 55–92. DOI : `10.1016/0890-5401(91)90052-4` (cf. p. 63).

Montague, Richard (1969). "On the nature of certain philosophical entities". In : *The Monist* 53.2, p. 159–194 (cf. p. 22).

— (1970a). "English as a formal language". In : *Linguaggi nella Società e nella Tecnica*. Milan : Edizioni di Comunità, p. 189–224 (cf. p. 1, 8, 11, 21, 46).

— (1970b). "Universal grammar". In : *Theoria* 36, p. 373–398 (cf. p. 2, 8, 21).

— (1973a). "Pragmatics and intensional logic". In : *Semantics of natural language*. Springer, p. 142–168 (cf. p. 22).

— (1973b). "The Proper Treatment of Quantification in Ordinary English". In : *Approaches to Natural Language*. Sous la dir. de Jaakko Hintikka et Patrick Moravcsik Julius andSuppes. T. 49. Reidel : Dordrecht, p. 221–242 (cf. p. 8, 21, 23, 63).

— (1974). *Formal Philosophy : Selected papers of Montague, Richard*. Yale University Press (cf. p. 11, 22).

Moortgat, M. (1988). *Categorial Investigations : Logical and Linguistic Aspects of the Lambek Calculus*. Groningen-Amsterdam studies in semantics. Foris Publications. ISBN : 9789067653879 (cf. p. 2).

Moosdijk, Sara von de (2014). "Mining texts at the discourse level". Mém.de mast. Université de Lorraine (cf. p. 109).

Morrill, Glyn V (1994). *Type Logical Grammar : Categorial Logic of Signs*. research monograph : Kluwer Academic Publishers, Dordrecht (cf. p. 2).

Muller, Philippe, Stergos Afantenos et al. (2012). "Constrained decoding for text-level discourse parsing". In : *COLING-24th International Conference on Computational Linguistics* (cf. p. 111).

Muller, Philippe et Laurent Prévot (2008). "The rhetorical attachment of questions and answers". In : *Meaning, Intentions, and Argumentation*. Sous la dir. de Kepa Korta et Joana Garmendia. T. 186. (CSLI-LN) Center for the Study of Language and Information - Lecture Notes. http ://www.journals.uchicago.edu/ : University of Chicago Press, p. 1–17 (cf. p. 126, 127).

Musiol, Michel et Alain Trognon (1996). "L'accomplissement interactionnel du trouble schizophrénique". In : *Raisons Pratiques 7*, p. 179–209 (cf. p. 80, 87).

Muskens, Reinhard (1996). "Combining Montague semantics and discourse representation". In : *Linguistics and philosophy* 19.2, p. 143–186 (cf. p. 3, 26).

Newman, M. H. A. et A. M. Turing (1942). "A formal theorem in Church's theory of types". In : *Journal of Symbolic Logic* 7, p. 28–33. ISSN : 0022-4812 (print), 1943-5886 (electronic) (cf. p. 17).

Oepen, Stephan et al. (2002). "LinGO Redwoods. A Rich and Dynamic Treebank for HPSG". In : *In Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT 2002)*. Sozopol, Bulgaria (cf. p. 111).

Orosanu, Luiza et Denis Jouvet (2015). "Detection of sentence modality on French automatic speech-to-text transcriptions". In : *International Conference on Natural Language and Speech Processing*. Alger, Algeria. HAL archive ouverte : `hal-01184193` (cf. p. 117).

Padrovni, Stéphanie (2015). "Description de l'articulation des registres visuels, (eye-tracking) et verbaux dans le maintien de l'interaction schizophrénique". Thèse de doct. Université de Lorraine (cf. p. 82).

Parigot, Michel (1992). "$\lambda\mu$-calculus : An Algorithmic Interpretation of Classical Natural Deduction". In : *proceedings of the Fourth International Conference on Types Lambda Calculi and Applications*, p. 190–201 (cf. p. 35, 43).

Partee, Barbara Hall (1984). "Nominal and temporal anaphora". In : *Linguistics and philosophy* 7.3, p. 243–286 (cf. p. 25).

Partee, Barbara Hall et Herman L.W. Hendriks (1997). "Montague Grammar". In : *Handbook of Logic and Language*. Sous la dir. de Johan van Benthem et Alice ter Meulen. MIT Press. Chap. 1. ISBN : 0262220539 (cf. p. 2, 5).

Peano, Giuseppe (1889). *Arithmetices principia : nova methodo*. Fratres Bocca (cf. p. 14).

Penstein Rosé, Carolyn et al. (1995). "Discourse Processing of Dialogues with Multiple Threads". In : *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. Cambridge, Massachusetts, USA : Association for Computational Linguistics, 31–38. DOI : `10.3115/981658.981663`. Anthologie ACL : `P95-1005` (cf. p. 125).

Perrier, Guy et Bruno Guillaume (2013a). "FRIGRAM : a French Interaction Grammar". In : *ESSLLI 2013 - Workshop on High-level Methodologies for Grammar Engineering*. Sous la dir. de Denys Duchier et Yannick Parmentier. Université de Düsseldorf. Düsseldorf, Germany, p. 63–74. HAL archive ouverte : `hal-00920717` (cf. p. 123).

— (2013b). *Leopar : an Interaction Grammar Parser*. Sous la dir. de Denys Duchier et Yannick Parmentier. ESSLLI 2013 - Workshop on High-level Methodologies for Grammar Engineering. Université de Düsseldorf. HAL archive ouverte : `hal-00920728` (cf. p. 123).

Péry-Woodley, Marie-Paule et al. (2011). "La ressource ANNODIS, un corpus enrichi d'annotations discursives". In : *Traitement Automatique des Langues* 52.3, p. 71–101. HAL archive ouverte : `halshs-00935201` (cf. p. 111).

Petronis, Arturas (2004). "The origin of schizophrenia : genetic thesis, epigenetic antithesis, and resolving synthesis". In : *Biological psychiatry* 55.10, p. 965–970 (cf. p. 90).

Peyraube, Alain (2002). "L'évolution des structures grammaticales". In : *Langages*, p. 46–58 (cf. p. 13).

Plotkin, Gordon D et Matija Pretnar (2013). "Handling algebraic effects". In : *arXiv preprint arXiv :1312.1399* (cf. p. 61).

Plotkin, Gordon et Matija Pretnar (2009). "Handlers of algebraic effects". In : *Programming Languages and Systems*. Springer, p. 80–94. DOI : `10.1007/978-3-642-00590-9_7` (cf. p. 6).

Pogodalla, Sylvain (2004). "Computing Semantic Representation : Towards ACG Abstract Terms as Derivation Trees". In : *Seventh International Workshop on Tree Adjoining Grammar and Related Formalisms - TAG+7*. Colloque avec actes et comité de lecture. Internationale. Vancouver, BC, Canada, p. 64–71. HAL archive ouverte : `inria-00107768` (cf. p. 123).

— (2009). *Advances in Abstract Categorial Grammars : Language Theory and Linguistic Modeling. ESSLLI 2009 Lecture Notes, Part II*. working paper or preprint. HAL archive ouverte : `hal-00749297` (cf. p. 123).

Portner, Paul et Barbara H. Partee, éds. (2002). *Formal Semantics : The Essential Readings*. Blackwell Publishers (cf. p. 5, 140).

Qian, Sai (2009). "Identification of accessibility constraints in discourse". Mém.de mast. Université de Nancy 2 (cf. p. 46).

— (2014a). "Accessibility of Referents in Discourse Semantics". Thèse de doct. Université de Lorraine. HAL archive ouverte : `tel-01104091` (cf. p. 5, 46).

— (2014b). "Discourse referents accessibility in discourse semantics". Thèse de doct. Université de Lorraine (cf. p. 49, 53, 58, 60).

Qian, Sai et Maxime Amblard (2011). "Event in compositional dynamic semantics". In : *Logical Aspects of Computational Linguistics*. Springer. DOI : `10.1007/978-3-642-22221-4_15` (cf. p. 51).

— (2012). "Accessibility for Plurals in Continuation Semantics". In : *The Forth JSAI International Symposia on AI (isAI2012) - Proceedings of the Ninth International Workshop of Logic and Engineering of Natural Language Semantics 9 (LENLS 9)*.

978-4-915905-51-3 C3004 (JSAI). Myasaki, Japon, p. 52–65. HAL archive ouverte : hal-00762203 (cf. p. 50).

Qian, Sai, Philippe de Groote et Maxime Amblard (2016). "Modal Subordination in Type Theoretic Dynamic Logic". In : *Linguistic Issues in Language Technology*. Modes of Modality in NLP 14, p. 54. HAL archive ouverte : hal-01370557 (cf. p. 5, 58, 60).

Quine, Willard Van Orman (1960). *Word and Object*. Cambridge, Mass, MIT Press (cf. p. 89).

Ranta, Aarne (2004). "Computational Semantics in Type Theory". In : *Mathematics and Social Sciences* 165, p. 31–57 (cf. p. 2).

Rebuschi, Manuel (2015). *Modélisation et rationalité dans l'analyse linguistique de conversations pathologiques*. Rencontres doctorales internationales en philosophie des sciences (cf. p. 122).

Rebuschi, Manuel, Maxime Amblard et Michel Musiol (2012). "Schizophrénie, logicité et compréhension en première personne". In : *L'Évolution psychiatrique* to appear (cf. p. 78).

— (2014). "Using SDRT to analyze pathological conversations. Logicality, rationality and pragmatic deviances". In : *Interdisciplinary Works in Logic, Epistemology, Psychology and Linguistics : Dialogue, Rationality, and Formalism*. Logic, Argumentation & Reasoning. Springer, p. 343–368. ISBN : 978-3-319-03043-2. HAL archive ouverte : hal-00910725 (cf. p. 90).

Reichenbach, H (1957). "The philosophy of space and time". In : (cf. p. 16, 24).

Reichenbach, Hans (1980). "Elements of symbolic logic". In : (cf. p. 16).

Retoré, Christian (2004). "A description of the non-sequential execution of Petri nets in partially commutative linear logic". In : *Logic Colloquium 99* Lecture Notes in Logic, p. 152–181 (cf. p. 33, 34).

Roberts, Craige (1987). "Modal subordination, anaphora, and distributivity". Thèse de doct. University of Massachusetts Amherst (cf. p. 55).

— (1989). "Modal subordination and pronominal anaphora in discourse". In : *Linguistics and philosophy* 12.6, p. 683–721 (cf. p. 54, 55).

Rooth, Mats (1987). "Noun Phrase Interpretation in Montague Grammar, File Change Semantics, and Situation Semantics". In : *Generalized Quantifiers*. Sous la dir. de Peter Gärdenfors. T. 31. Studies in Linguistics and Philosophy. Springer Netherlands, p. 237–268. ISBN : 978-1-55608-018-0. DOI : 10.1007/978-94-009-3381-1_9 (cf. p. 26).

Roze, Charlotte, Laurence Danlos et Philippe Muller (2012). "Lexconn : A french lexicon of discourse connectives". In : *Discours, Multidisciplinary Perspectives on Signalling Text Organisation* 10, (on line) (cf. p. 125).

Ruet, Paul et François Fages (1998). "Concurrent constraint programming and non-commutative logic". In : *Computer Science Logic*. Sous la dir. de Mogens Nielsen et Wolfgang Thomas. T. 1414. Lecture Notes in Computer Science. Springer Berlin Heidelberg, p. 406–423. ISBN : 978-3-540-64570-2. DOI : 10.1007/BFb0028028 (cf. p. 34).

Russell, Bertrand (1905). "On denoting". In : *Mind* 14.56, p. 479–493 (cf. p. 15).

— (1921). *Analysis of Mind*. George Allen & Unwin (cf. p. 15).

Sagot, Benoit (2010). "The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French". In : *7th international conference on Language Resources and Evaluation (LREC 2010)*. Valletta, Malta. HAL archive ouverte : `inria-00521242` (cf. p. 99).

Saurer, Werner (1993). "A natural deduction system for discourse representation theory". In : *Journal of Philosophical Logic* 22.3, p. 249–302 (cf. p. 25).

Schmid, Helmut (1994). "Probabilistic Part-of-Speech Tagging Using Decision Trees". In : *Proceedings of International Conference on New Methods in Language Processing* (cf. p. 99).

Sells, Peter (1985). *Restrictive and non-restrictive modification*. T. 28. Center for the Study of Language et Information, Stanford University (cf. p. 55).

Shan, Chung-chieh (2004). "Delimited continuations in natural language : Quantification and polarity sensitivity". In : *arXiv preprint cs/0404006* (cf. p. 46).

Shannon, Claude E (1949). "Communication in the presence of noise". In : *Proceedings of the IRE* 37.1, p. 10–21 (cf. p. 19).

Soricut, Radu et Daniel Marcu (2003). "Sentence Level Discourse Parsing Using Syntactic and Lexical Information". In : *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*. NAACL '03. Edmonton, Canada : Association for Computational Linguistics, p. 149–156. DOI : `10.3115/1073445.1073475` (cf. p. 111).

Sporleder, Caroline et Alex Lascarides (2008). "Using automatically labelled examples to classify rhetorical relations : an assessment". In : *Natural Language Engineering* 14 (03), p. 369–416 (cf. p. 111).

Stabler, Edward (1997). "Derivational Minimalism". In : *Logical Aspect of Computational Linguistic* 1328. Sous la dir. de Springer-Verlag (cf. p. 29, 30).

Stark, Lawrence (1983). "Abnormal patterns of normal eye movements in schizophrenia." In : *Schizophrenia Bulletin* 9.1, p. 55–72 (cf. p. 82).

Stone, Matthew et Daniel Hardt (1999). "Dynamic discourse referents for tense and modals". In : *Computing meaning*. Springer, p. 301–319 (cf. p. 24).

Strachey, Christopher et Christopher P. Wadsworth (1974). *Continuations : A mathematical semantics for handling full jumps*. Rapp. tech. Oxford University Computing Laboratory, Programming Research Group (Oxford) (cf. p. 46).

Struik, D.J. (1969). *A Source Book in Mathematics, 1200-1800*. Source Books in the History of the Sciences. Harvard University Press. ISBN : 9780674823556 (cf. p. 14).

Tarski, Alfred (1933). "The Concept of Truth in Formalized Languages". In : *Nakładem Towarzystwa Naukowego Warszawskiego (Polish original version)* German translation with an added postscript, Tarski 1935 ; English version of the German text, Tarski 1983b (cf. p. 20).

— (1944). "The semantic conception of truth : and the foundations of semantics". In : *Philosophy and phenomenological research* 4.3, p. 341–376 (cf. p. 11, 20).

— (1956). "The concept of truth in formalized languages". In : *Logic, semantics, metamathematics* 2, p. 152–278 (cf. p. 20).

Tarski, Alfred, Joseph Henry Woodger et John Corcoran (1956). *Logic, semantics, metamathematics : papers from 1923 to 1938*. Clarendon Press Oxford (cf. p. 20).

Tiv, Stéphan (2016). "Modelling Dialogues in a Dynamics Framework : formal integration and evaluation". Mém.de mast. Université de Lorraine (cf. p. 125, 128).

Turing, A. M. (1937). "Computability and λ-definability". In : *Journal of Symbolic Logic* 2, p. 153–163. ISSN : 0022-4812 (print), 1943-5886 (electronic) (cf. p. 17, 18).

— (1950). "Computing Machinery and Intelligence". In : *Mind* 59.236, p. 433–460. ISSN : 0026-4423 (print), 1460-2113 (electronic) (cf. p. 17).

Urieli, Assaf et Ludovic Tanguy (2013). "L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane". In : *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*. Les Sables d'Olonne, France, p. 188–201 (cf. p. 116).

van Benthem, Johan (1986). *Essays in logical semantics*. Studies in linguistics and philosophy 29. D. Reidel Pub. Co. ISBN : 9789027720924 (cf. p. 2).

— (1988). "The semantics of variety in categorial grammar". In : *Categorial grammar* 25, p. 37–55 (cf. p. 2, 29).

— (1991). "Quantifiers in The World of Types". In : *Quantification in the Netherlands*. Sous la dir. de Jaap van der Does et Jan van Eyck. Institute for Language, Logic et InformationUniversity of Amsterdam (cf. p. 25, 26).

— (1995). *Language in action : Categories, lambdas, and dynamic logic*. The MIT Press (cf. p. 3).

van Eijck, Jan (1995). "Presuppositions and dynamic logic". In : *Papers from the Second CSLI Workshop on Logic, Language and Computation*. Sous la dir. de Makoto Kanazawa, Christopher Piñon et Henriette de Swart. Stanford : CSLI Publications (cf. p. 25).

Van Leusen, Noor et Reinhard Muskens (2002). *Construction by description in discourse representation*. Institute for Logic, Language et Computation (ILLC), University of Amsterdam (cf. p. 25).

Van Rooij, Robert (2005). "A modal analysis of presupposition and modal subordination". In : *Journal of Semantics* 22.3, p. 281–305 (cf. p. 55).

von Fintel, Kai (2006). "Modality and language". In : *Encyclopedia of philosophy - second edition*. Sous la dir. de Donald M. Borchert. Detroit : Macmillan Reference USA (cf. p. 54).

von Heusinger, K. et A. Meulen (2013). *Meaning and the Dynamics of Interpretation : Selected Papers of Kamp, Hans*. Current Research in the Semantics / Pragmatics Interface. Brill. ISBN : 9789004252882 (cf. p. 121).

Vu, Anh-Duc (2016). "Text mining at discourse level". Mém.de mast. Université de Lorraine (cf. p. 109).

Wadler, Philip (1992). "The essence of functional programming". In : *Proceedings of the 19th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*. ACM, p. 1–14 (cf. p. 63).

Wechsler, D. (1958). *The Measurement and Appraisal of Adult Intelligence*. The Measurement and Appraisal of Adult Intelligence. Williams & Wilkins (cf. p. 81).

Wellner, Ben et al. (2009). "Classification of discourse coherence relations : An exploratory study using multiple knowledge sources". In : *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, p. 117–125 (cf. p. 111).

Whitehead, Alfred North et Bertrand Russell (1912). *Principia mathematica*. T. 2. University Press (cf. p. 15).

Wolf, Florian et al. (2004). "Discourse Graphbank". In : *Philadelphia, PA : Linguistic Data Consortium* (cf. p. 111).

Woods, Steven Paul et al. (2006). "The California Verbal Learning Test – second edition : Test-retest reliability, practice effects, and reliable change indices for the standard and alternate forms". In : *Archives of Clinical Neuropsychology* 21.5, p. 413–420. ISSN : 0887-6177. DOI : 10.1016/j.acn.2006.06.002 (cf. p. 81).

Zeevat, Henk (1989). "A compositional approach to discourse representation theory". In : *Linguistics and Philosophy* 12.1, p. 95–131 (cf. p. 26).