Habilitation à diriger des recherches

soutenue à

L'Université Charles de Gaulle - Lille 3

Discipline : Informatique

par

Isabelle TELLIER

Modéliser l'acquisition de la syntaxe du langage naturel via l'hypothèse de la primauté du sens

---

Date de soutenance : jeudi 8 décembre 2005

Devant un jury composé de :

Pieter Adriaans	Université d'Amsterdam	Rapporteur
Alexandre Dikovsky	Université de Nantes	Rapporteur
Rémi Gilleron	Université Lille 3	Directeur
Laurent Miclet	ENSSAT Lannion	Examinateur
Philip Miller	Université de Lille 3	Examinateur
Christian Rétoré	Université de Bordeaux	Rapporteur

Résumé :

L'objet de ce travail est la modélisation informatique de la capacité d'apprentissage de la syntaxe de leur langue naturelle par les enfants. Une synthèse des connaissances psycho-linguistiques sur la question est donc tout d'abord proposée. Le point de vue adopté pour la modélisation accorde une place privilégiée à la sémantique, qui est supposée acquise avant la syntaxe. Le Principe de compositionnalité, éventuellement adapté, est mis à contribution pour formaliser les liens entre syntaxe et sémantique, et le modèle d'apprentissage "à la limite" par exemples positifs de Gold est choisi pour régir les conditions de l'apprentissage.

Nous présentons dans ce contexte divers résultats d'apprenabilité de classes de grammaires catégorielles à partir de divers types de données qui véhiculent des informations sémantiques. Nous montrons que, dans tous les cas, la sémantique contribue à spécifier les *structures* sous-jacentes aux énoncés, et à réduire ainsi l'espace de recherche des algorithmes d'apprentissage.

Mots clés :

grammaires catégorielles, interface syntaxe-sémantique, Principe de compositionnalité, inférence grammaticale, modèle de Gold

Abstract :

This work deals with the computational modeling of natural language syntax learning by children. A synthesis of contemporary psycho-linguistic knowledge on the subject is first proposed. The focus is put on semantics, which is supposed to be first acquired. The Principle of compositionality, sometimes adapted, is adopted to formalize the links between syntax and semantics, and Gold's model of learnability "in the limit" from positive examples is chosen to specify the conditions where learning takes place.

We show, under these conditions, various learnability results for classes of categorial grammars, for various kinds of input data carrying semantic information. We show that, in each case, semantic information contribute to specify the underlying *structures* of utterances, and so to reduce the search space of the learning algorithms.

Key words :

categorial grammars, syntax-semantic interface, Principle of compositionality, grammatical inference, Gold's model

Remerciement

Les recherches exposées ici sont le fruit d'une dizaine d'années de travail et de vie ; il va donc être difficile de n'oublier aucun de ceux qui y ont contribué d'une manière ou d'une autre.

Mes premiers remerciements vont à Rémi Gilleron, auquel j'associerai volontiers François Denis, les deux "papys fondateurs" de l'équipe qui m'a accueillie à mon arrivée à Lille3 en 1996, et qui ne s'appelait pas encore Grappa. J'admire beaucoup la façon tranquille et exigeante dont ils ont monté, fait grandir progressivement et évoluer cette équipe. Et merci à Rémi pour toutes les gaffes qu'il m'a signalées dans la première version de ce document.

La thèse de Pieter Adriaans a été le premier document que François et Rémi m'ont mis entre les mains en 1996 pour commencer à travailler avec eux. On peut dire que les recherches que j'ai menées depuis sont les héritières directes de cette lecture. J'ai depuis croisé plusieurs fois Pieter dans des conférences ; je suis donc évidemment très heureuse qu'il ait accepté de participer à ce jury.

Je connais le nom de Christian Rétoré depuis quasiment aussi longtemps que je connais les grammaires catégorielles -c'est dire ! J'ai toujours apprécié ses idées et son enthousisame, et ai beaucoup bénéficié de ses initiatives pour animer la communauté de ceux qui s'y intéressent -en particulier l'ARC Gracq de l'Inria. Je ne pouvais pas imaginer un jury sans lui.

Alexandre Dikovsky est du genre passionné, sa passion est terriblement contagieuse. Je l'ai surpris un week-end, sur le banc d'un supermarché, en train de prendre des notes sur un article de Pinker. Ses idées sur la modélisation de l'apprentissage du langage des enfants sont parmi les plus originales et les plus stimulantes que j'aie lues. Il est particulièrement le bienvenu dans ce jury.

Le "Miller et Torris" a été une de mes premières références bibliographiques, quand j'ai découvert, il y a pas mal années déjà, les grammaires catégorielles. J'ai quelquefois croisé Philip Miller dans les couloirs du labyrinthe de Lille3, sans avoir jamais pris beaucoup le temps de discuter avec lui. Je suis très honorée qu'il soit présent dans ce jury, pour rattraper un peu de ce temps perdu.

Laurent Miclet est un des piliers de la communauté de l'apprentissage automatique en France. Je lui dois plein de discussions intéressantes et quelques fous rires. Je le sais très occupé, je suis ravie et très reconnaissante qu'il puisse venir participer à ce jury.

Je pense aussi à cette occasion à mes parents et à mes frères, Pascal et Dominique, et à leur famille, qui habitent trop loin pour profiter des petits fours. J'y joins tous mes amis de la PU (ils se reconnaîtront), mes collègues de l'Idist, de philo, de psycho, de Lettres modernes, de Lille1 ou du CNU, tous ceux de l'ARC Gracq, particulièrement Annie et Denis. Je pense à tous les amis d'avant mon arrivée à Lille dont je n'ai pas perdu la trace (Gilles, Michaël, Isabelle, Grégoire,

Cécile, Claude, Françoise, Pascal...), et à ceux que j'ai connus depuis qui en ont commencé de nouvelles (Catherine, Marion, Sophie, Thierry), tous victimes de mes innombrables squats parisiens ou provinciaux.

Mais je dédie principalement ce document à feue la "Petite Maison des informaticiens" -expression à prendre bien sûr dans son acception métonymique, pour désigner tous ceux qui ont occupé ses murs à un moment ou un autre. J'ai eu beaucoup de chance de vivre et de travailler avec eux. Pour Marc, donc, et tous les gros mots d'informatique qu'il nous a appris, pour Dominique et ses blagues idiotes (je suis injuste, il n'y a pas que les siennes), pour les bricolages d'Alain, les bonbons de Fabien, les critiques de Joaquim, les silences éloquents de Rémi. Pour les derniers arrivés qui se renforcent mutuellement (Philippe et Rémi-bis), pour les papys-bis toujours là quand il le faut -surtout pour les pots- (Jean, Alain-bis et Luc) et les rires sonores d'Anne. Pour tous les ex-thésards d'hier (Aurélien Daniela, Julien), les ex-Ater d'aujourd'hui (Benjamin) et les futurs ex-étudiants de demain (Laurent, Patrick, Florent, Fabien-bis) qui raconteront un jour qu'ils en ont été, aussi. Pour tous les séminaires, les débats, les bières et les galettes que nous avons partagés. Pour tous les cafés que nous avons bus ensemble, et tous ceux que j'ai renversés sur eux dans ses murs (et ailleurs). Pour les stagiaires qui y ont découvert la vie de laboratoire, pour les étudiants qui y ont mendié des points, pour les intrus qui avaient perdu leur chemin, et pour un cosmonaute belge.

Les documents universitaires sont souvent agrémentés de petites citations ayant un plus ou moins lointain rapport avec leur sujet. Il se trouve que j'en ai un stock collecté depuis quelques années au gré de mes lectures, et que, ne sachant choisir, je préfère les livrer toutes. Elles constituent comme un portrait chinois, en plus ludique, des recherches "sérieuses" qui suivent.

****

Peu à peu, je fis une découverte encore plus importante. Ces gens possédaient un moyen de communiquer leurs pensées et leurs sentiments par des sons. [...] En m'appliquant beaucoup et après être resté pendant plusieurs changements de lune dans ma cabane, je découvris les noms qu'ils donnaient à quelques uns des objets les plus familiers. [...] Je ne peux décrire ma joie lorsque j'appris le sens propre de chacun de ces sons et que je pus les prononcer.

Mary Shelley, *Frankenstein*, chapitre 12

Je dis souvent "Mmmm..." J'ai signalé le fait à Karla. Elle a dit que ça venait de mon unité centrale. "C'est la mise en attente, le temps que tu assembles les données dans ta tête." Et je dis souvent "on dirait". Pour ce cas, a dit Karla, il n'existe pas d'explication opératoire. [...] Je crois que je vais essayer un Chercher-Remplacer mental pour éliminer ces deux maudits mots. Essayer de me déboguer.

Douglas Coupland, *Microserf*

Les anciens Egyptiens croyaient que le siège de l'âme résidait dans la langue : la langue était un gouvernail ou une godille qui permettait à l'homme de diriger sa course dans le monde.

Bruce Chatwin, *Le chant des pistes*

Knowing a language, then, is knowing how to translate mentalese into strings of words and vice-versa.

Stephen Pinker, *The Language Instinct*

Pourquoi ces deux approches devraient-elles s'exclure mutuellement ? Pourquoi l'acquisition du langage ne pourrait-elle pas impliquer à la fois le mécanisme formel de Turing pour la syntaxe et une capacité générale d'apprentissage pour attacher un sens aux symboles ? Je ne voit pas de raison logique qui interdirait que les choses se passent ainsi.

John Casti, *Un savant dîner*

Quel homme s'est imaginé que ce sinistre automate pourrait m'émouvoir à l'aide d'on ne sait quels paradoxes inscrits dans des feuilles de métal ! Depuis quand Dieu permit-il aux machines de prendre la parole ?

Villiers de l'Isle-Adam, *L'ève future*, livre VI, chapitre X

La parole possède une fonction émotive inouïe qui nous permet de pleurer pour un événement survenu il y a vingt ans, ou d'espérer une situation qui ne se présentera que dans dix ans. Le sens, introduisant l'absent dans le présent, peut plonger, par lui, dans un passé dont on ne voit pas les limites, pas plus qu'on n'en discerne à l'avenir.

Boris Cyrulnik, *La naissance du sens*

Bien que l'histoire des écritures n'ait jamais été sa spécialité, par ses anciennes études il se souvient plus ou moins de Champollion et de sa façon de déchiffrer les hiéroglyphes égyptiens, de Grotefend avec les pierres perses et les écritures cunéiformes [...]. Dans tous les cas les chercheurs disposaient de vestiges polyglottes [...]. Mais lui ici, que peut-il faire avec l'écriture inconnue d'une langue inconnue, sans aucune aide extérieure ? De quelle hypothèse partir, compiler quoi avec quoi, sans référence, tout au moins pour le moment : quelle ligne de caractères rattacher à quel mot et quel sens attribuer à n'importe quel mot ?

Ferenc Karinthy, *Epépé*, p.45-46

Il est de fait que l'on est pris dans un cercle vicieux : pour rendre compte du sens, il faudrait que je puisse le formaliser. Mais si je fais ainsi, il devient information et cesse d'être sens. Alors existe-t-il ? c'est là que l'on retrouve les présupposés "religieux" : je crois à l'existence du SENS.

Jacques Arsac, *La science informatique*

Si un certain mot était attribué tantôt à une chose et tantôt à une autre, ou encore si la même chose était appelée tantôt d'un nom et tantôt d'un autre, sans qu'il y eût aucune règle à laquelle les phénomènes fussent déjà soumis eux-mêmes, aucune synthèse empirique de l'imagination ne pourrait avoir lieu.

Emmanuel Kant, *ref. inconnue*

Le difficile [...] c'est d'établir une équation qui tienne compte de toutes les conditions du problème. Le reste n'est plus qu'une question d'arithmétique, et n'exige que la connaissance des quatres règles.

Jules Verne, *Autour de la lune*, chapitre IV

[...] il avait étudié les merveilles du corps humain et essayé de sonder le processus par lequel la Nature emprunte toutes ces précieuses influences à la terre, à l'air et aux régions de l'esprit pour créer et maintenir l'homme, son chef d'oeuvre. Ce dernier champ d'étude, cependant, Ayhmer l'avait laissé de côté, obligé de reconnaître malgré lui le fait auquel se heurtent tôt ou tard tous les chercheurs, que notre puissante Mère créatrice, qui se plait en apparence à oeuvrer au plein soleil de la plus claire lumière, prend cependant le plus grand soin de garder son secret, et, en dépit d'un semblant de franchise, ne nous livre que les résultats. Elle nous permet, il est vrai, de gâter, mais rarement d'améliorer, et comme le possesseur jaloux d'un brevet, jamais en aucune façon de créer.

Nathaniael Hawthorne, *La marque sur le visage*

The theory of mental models suggests that what children have to acquire are the truth conditions of expressions - more accurately the contribution that expressions make to the truth conditions of sentences.

Philip N. Johnson-Laird, *Mental Models*, chapitre 11

Notre but ultime est de concevoir des programmes qui apprennent par expérience, comme le font effectivement les êtres humains.

attribué à MC Carthy par Daniel Crevier, *A la recherche de l'intelligence artificielle*

Ecrire, c'est déjà traduire dans une autre langue.

Marina Svetaeva, *ref. inconnue*

Ludmilla m'a parlé de son travail en équipe sur la modélisation du processus d'apprentissage, cherchant à programmer l'accumulation des connaissances apportées par l'expérience, le dialogue avec d'autres apprenants et leur imitation, ainsi que l'acquisition de règles, faisant appel à la simulation de jeux enfantins [...]. L'entreprise m'a paru assez désespérée, trop de variables, mais je me suis poliment borné à des commentaires encourageants.

David Lodge, *Pensées secrètes*, p.261-262

That is, before children have learned syntax, they know the meaning of many words, and they might be able to make good guesses as to what their parents are saying, based on their knowledge of how the referents of these words typically act [...].

Stephen Pinker, *Language*, chapitre "Language Acquisition"

D'où vient que l'on parle ? Que la viande s'exprime ?

Valère Novarina, *Le drame de la vie*

A l'audition de cette phrase, il se passa dans la cervelle du comte, habitée par le *moi* d'Octave, un très singulier phénomène : les sons étrangers au Parisien, suivant les replis d'une vieille oreille slave, arrivèrent à l'endroit habituel où l'âme d'Olaf les accueillait pour les traduire en pensées, et y évoquèrent une sorte de mémoire physique ; leur sens apparent confusément à Octave ; des mots enfouis dans les circonvolutions cérébrales, au fond des tiroirs secrets du souvenir, se présentèrent en bourdonnant tout prêts à la réplique ; mais ces réminiscences vagues, n'étant pas mises en communication avec l'esprit, se dissipèrent bientôt, et tout redevint opaque.

Théophile Gautier, *Avatar*

Ah ! Mon crâne est ouvert ! Je souffre horriblement ! - Que cherchent-ils dans ma tête ?... la pensée peut-être ?... C'est au nom de la science que les barabares nous hachent, nous dépècent et nous fouillent !...

Claude Vignon, *Les morts se vengent*

- How can you talk if you haven't got a brain ?
- I don't know, but some people without brains do a lot of talking, don't they ?
- Yes, I guess you're right.

dans le film *Le magicien d'Oz*, 33ème mn

Rien de si simple, comme on voit, que la mécanique de notre éducation ! Tout se réduit à des sons ou à des mots, qui de la bouche de l'un passent par l'oreille de l'autre dans le cerveau, qui reçoit en même temps par les yeux la figure des corps dont ces mots sont les signes arbitraires.

Julien Offray de la Mettrie, *L'homme machine*

Table des matières

Modéliser l'acquisition de la syntaxe du langage naturel via l'hypothèse de la primauté du sens

Chapitre 1

Introduction

Même si le terme d'"Intelligence Artificielle" est né officiellement en 1956, lors de la fameuse conférence de Dartmouth (Crevier, 1999), c'est bien sûr à Turing qu'il faut attribuer la paternité du projet de construire une machine logique "intelligente" (Hodges, 1983). Dans l'article intitulé "Computing Machinery and Intelligence" qu'il fait paraître en 1950 dans la revue "Mind" (Turing, 1950), il présente ce qui est connu depuis comme le "test de Turing". La proposition est habile : plutôt que de chercher à définir ce qu'est l'intelligence, Turing se concentre sur ses effets observables dans les interactions entre humains, en particulier dans leurs conversations spontanées. Et il poursuit en remarquant que ce que chacun attribue spontanément à un semblable avec qui il dialogue, il n'a aucune raison de le dénier à un dispositif artificiel qui, à distance, en est indistinguable. Ce test semble ainsi faire de la *compétence langagière commune*, mise en scène dans un dialogue, un critère indirect de l'intelligence.

Cet article fondateur est aussi intéressant à un autre titre : il se termine par la vision, très audacieuse au vu des connaissances de l'époque, d'un programme qui "simulerait le cerveau d'un enfant". Soumis à une "éducation appropriée", un tel programme pourrait acquérir des capacités qu'il aurait été beaucoup plus difficile de lui donner explicitement dès le départ. Le projet de l'apprentissage automatique était déjà là, en germe.

Mais Turing lui-même ne fait pas le lien entre les deux parties de son article : l'apprentissage du langage ne figure pas au programme éducatif de son "bébé artificiel". Pourtant, l'acquisition de leur langue maternelle par les enfants est une étape cruciale de leur développement ; elle ne cesse, depuis des années, d'étonner les psychologues et les linguistes (Piatelli-Palmarini, 1979; de Boysson-Bardie, 1999). Elle interpelle aussi les éthologues ; certains d'entre eux dépensent beaucoup d'efforts pour enseigner à des animaux des bribes de langages (Dessalles, 2000). La maîtrise d'une langue semble bien une capacité spécifiquement humaine, qui plus est acquise très tôt.

Encore aujourd'hui, alors que l'ordinateur dépasse l'homme le plus performant dans beaucoup de tâches intellectuelles prestigieuses (par exemple pour jouer aux échecs), il s'avère incapable de passer le test de Turing, et d'accéder aux compétences langagières de "l'homme de la rue". Le domaine du "traitement automatique des langues" a pourtant connu un développement considérable ces dernières années. L'évolution des capacités de stockage et de traitement des machines a notamment permis de faire émerger de nouvelles approches à base de corpus, inenvisageables auparavant. Mais le problème est loin d'être résolu pour autant.

La maîtrise du langage est donc typiquement l'une de ces capacités humaines, évoquées par Turing, qu'il est très difficile de programmer directement, et qu'il faudrait plutôt approcher par le biais de son apprentissage. Les travaux introduits ici portent sur la *modélisation de la capacité d'acquisition du langage par les enfants.* Ils ne prétendent évidemment pas faire le tour de la question, tant le problème est complexe et les façons de l'aborder variées. Tâchons donc tout d'abord de préciser les termes de ce projet.

La notion de modélisation employée ici est celle qui prévaut en informatique. Pour nous, l'informatique est *la science des traitements effectifs applicables à des données discrètes.* Modéliser une situation ou un processus, dans notre sens, cela signifie donc d'une part identifier les données représentatives de ce processus et les coder dans des *structures de données*, d'autre part transcrire les modifications subies par ces données dans des *algorithmes.* Cette démarche est représentée dans la Figure 1.1, inspirée d'un schéma de Jacques Arsac, un des pionniers de l'informatique en France (Arsac, 1987; Arsac & Vauthier, 1989).

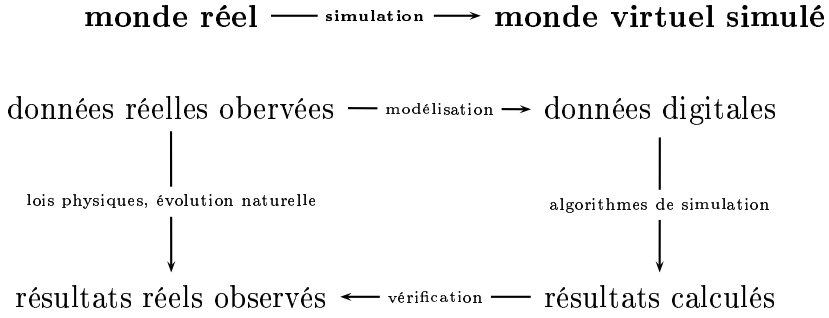

FIG. 1.1 – démarche de modélisation en informatique

Mais, même à l'intérieur de ce cadre, les approches peuvent être très diverses : celles que nous privilégierons, pour aborder aussi bien le langage que l'apprentissage, sont à base de modèles *symboliques* ou structurels, plutôt que statistiques, garantissant la possibilité de *preuves formelles* tout en autorisant aussi, autant que possible, des *validations empiriques.*

La pertinence d'un modèle informatique ou formel pour décrire un phénomène naturel est une question difficile (Dupuy, 1994; Brent, 1996), qui est au coeur du projet des *sciences cognitives* (Gardner, 1993). Nous ne prétendons pas y apporter de solution neuve. L'hypothèse *fonctionnaliste*, selon laquelle la notion de traitement de l'information peut être envisagée indépendamment du substrat biologique qui la rend possible, est évidemment sous-jacente. Mais il est tout aussi évident que ce n'est pas parce que les entrées et sorties d'un programme sont superficiellement comparables à celles observées dans la nature que les processus intermédiaires le sont aussi.

Reste à préciser la nature même du processus à modéliser. "L'apprentissage du langage" en général est trop complexe pour être abordé dans son intégralité. C'est sur une toute petite partie de ce processus que nous focaliserons nos efforts : l'acquisition de la *syntaxe*. Mais cela ne signifie pas pour autant que nous considérons la syntaxe indépendamment de tout autre niveau d'analyse. Nos travaux, en effet, partent de l'hypothèse que, dans le langage, c'est surtout la *sémantique* qui est essentielle. Une langue ne se réduit pas à un corpus, aussi significatif soit-il. Elle ne se limite pas non plus à un jeu de manipulations de symboles. Elle permet surtout de dire des choses sur le monde, de véhiculer du *sens*. Pour un humain, un abîme sépare une langue maîtrisée d'une langue inconnue, un texte familier d'une série de hiéroglyphes[1]. Si on veut modéliser cette propriété, il faut en quelque sorte "faire entrer le monde tel que l'homme le voit" dans la mémoire des ordinateurs[2]. Nous estimons que *l'acquisition sémantique est un préalable nécessaire à l'acquisition syntaxique*. Même si nos travaux portent plutôt sur la syntaxe, nous serons donc amenée à faire des hypothèses sur la nature de ces connaissances sémantiques préalables, et sur les liens qui relient la syntaxe et la sémantique.

Ce document est conçu comme une introduction très générale aux travaux particuliers que nous avons menés ces dernières années, et dont nous avons extrait six articles, reproduits à la suite. La partie 2 porte sur le domaine à modéliser : le langage naturel, ses particularités, et la façon dont les enfants l'apprennent. Cette partie relève donc principalement de la linguistique et de la psychologie. Les deux parties suivantes abordent respectivement comment on peut modéliser le langage (partie 3) et l'apprentissage (partie 4) en informatique. Ce n'est que dans la partie 5 que sont présentées nos contributions personnelles.

L'ambition principale de ces travaux est certes de simuler l'acquisition humaine. Mais ils peuvent être applicables à d'autres domaines, quitte à oublier un peu le critère de crédibilité psycholinguistique : acquisition automatique de lexiques, in-

[1] à moins bien sûr qu'il ne soit égyptologue !

[2] dans le "test de Turing", les compétences linguistiques ne sont pas explicitement mises en avant ; c'est que le langage y est surtout envisagé comme le moyen d'aborder tous les sujets possibles, et donc d'empêcher le programme qui y est soumis d'être spécialisé dans un domaine trop précis

férence de grammaires formelles pour éviter d'avoir à les écrire soi-même, apprentissage de patterns (grammaires locales) pour l'extraction d'information. Il pourra donc nous arriver d'évoquer ces domaines, au fur et à mesure que nous avancerons.

Nos contributions personnelles reposent, la plupart du temps, sur l'assemblage de modèles existants (modèles du langage/modèles d'apprentissage), la définition de nouvelles façons de les faire intéragir, plus que sur l'introduction ou l'affinement d'un modèle particulier. Il nous a donc semblé souhaitable, dans ce document, de présenter les différents candidats au statut de modèle que nous avons envisagés. Ce texte évoque donc beaucoup de domaines divers sans forcément les approfondir ; il s'attache plutôt à argumenter chacun des choix effectués, les alternatives possibles, leurs limites et leurs conséquences. Les aspects techniques (largement présents dans les articles) seront, eux, traités "au fil de la plume" avec le minimum de notation mathématique possible. Sans doute ce texte ressemble-t-il, souvent, plus à une réflexion méthodologique et épistémologique qu'à un exposé scientifique précis. Il requiert très peu de connaissances préalables, sinon les bases de la théories des langages formels, de la logique du premier ordre et du λ-calcul.

Enfin, les articles rassemblés à la fin du document se veulent chacun représentatif d'une approche propre (pas plus de deux pour chacune des 3 sous-parties de la partie 5). Nous n'avons choisi que des articles déjà publiés dans des actes de conférences, plutôt que des versions de synthèse qui auraient été plus complètes mais ont seulement été soumises.

Chapitre 2

Le langage et son apprentissage

Tous les groupes humains découverts de par le monde pratiquent au moins une langue. On en dénombre environ 5 000 différentes, dont beaucoup sont actuellement en voie de disparition. Chaque être humain normalement constitué et inséré depuis sa naissance dans un groupe social est capable, vers l'âge de 5 ans, de tenir une conversation courante dans sa langue maternelle -même si l'acquisition du vocabulaire se poursuit tout au long de la vie.

Les hommes passent une proportion considérable de leur temps de veille à parler. La fonction de cette activité (Jakobson en proposait 6, parmi lesquelles la fonction d'énonciation n'est qu'une parmi d'autres) et les avantages évolutifs qu'elle apporte à l'espèce n'ont rien d'évident à caractériser (Dessalles, 2000).

Maîtriser une langue requiert des aptitudes multiples et variées, que les linguistes ont progressivement mises à jour. Cette section est consacrée aux propriétés fondamentales des langues dites "naturelles" et aux conditions de leur apprentissage.

2.1 Le langage

Le langage, c'est cette "faculté de langue" universellement distribuée dans l'espèce humaine, que les linguistes essaient de caractériser à l'aide *d'universaux*. Nous présentons ici les propriétés du langage qui seront la base de notre approche, en insistant sur les liens entre les différents niveaux d'analyse auxquels il se prête.

2.1.1 Spécificités des langues naturelles

Ce qui fait la spécificité des langues humaines, dîtes encore "langues naturelles", par rapport aux autres modes de communication symboliques identifiés dans la nature (comme les cris d'animaux ou le vol des abeilles), ou par rapport aux

langues formelles inventées par les logiciens et les informaticiens est une question difficile qui a donné lieu à de multiples débats.

Plusieurs auteurs ont insisté sur le caractère combinatoire particulier des langues naturelles. Le linguiste André Martinet a proposé la notion de "double articulation" pour en rendre compte (Martinet, 1960). La première articulation, dans son système, est celle qui permet la combinaison "d'unités douées chacune d'une forme vocale et d'un sens" qu'il appelle des "monèmes" (d'autres auteurs parleront plutôt de *morphèmes*) pour constituer des énoncés complets. Il donne l'exemple de l'énoncé "j'ai mal à la tête", constitué des morphèmes "j", "ai", "mal", "à", "la" et "tête". La deuxième articulation décrit comment chaque morphème est lui-même décomposable en une succession d'unités phoniques élémentaires et dépourvues de sens, les *phonèmes*. Chacun de ces deux niveaux décrit une combinatoire *ouverte* (au sens où la liste des éléments qu'elle produit est potentiellement infinie) d'*éléments discrets*. Ce dispositif diffère fondamentalement des codages de type *analogique*, comme celui utilisé par les abeilles qui modulent l'amplitude et l'orientation de leur vol pour communiquer à leurs congénères l'emplacement d'une source de nourriture. Comme le remarque Pinker (Pinker, 1994), les deux systèmes les plus sophistiqués de transmission d'information sélectionnés par la nature, à savoir les langues naturelles et le code génétique, reposent tous les deux sur des unités discrètes.

Une grande partie de la tradition linguistique héritière de Chomsky (Chomsky, 1957; Chomsky, 1968) s'est employée à décrire par des règles précises la nature de la "première articulation" de Martinet, qui est bien sûr celle de la *syntaxe*. A l'origine de ce programme de recherche particulièrement fécond, se trouve l'hypothèse qu'*un nombre fini de règles syntaxiques* suffit à expliquer la production de toutes les phrases grammaticalement correctes possibles (évidemment en nombre infini). Cette propriété, caractéristique d'un processus *récursif*, est le point de départ des travaux sur les *grammaires formelles*, sur lesquelles nous reviendrons en partie 3.1.

Mais cette primauté accordée à la syntaxe tend à minimiser les différences entre les langues naturelles et les langues formelles des logiciens et des informaticiens. Elle éclipse ce qui fait l'incroyable pouvoir des langues : leur capacité à référer à ce qui leur est extérieur, à dire des choses sur le monde. Cette capacité opère également à deux niveaux. Saussure (de Saussure, 1916) caractérisait déjà le *signe linguistique* comme une association arbitraire entre un "signifiant" (une "image mentale, visuelle ou acoustique") et un "signifié" (un concept). La *sémantique lexicale* étudie, dans cette lignée, les sens qu'on peut associer aux morphèmes. Ils sont beaucoup moins simples qu'il n'y paraît au premier abord, parce qu'ils réfèrent, pour la plupart, à des *catégories* au sens de la psychologie (Lakoff, 1987; Nyckees, 1998) : qu'on pense à tous les individus fort dissemblables auxquels réfère un mot comme "chien". Les acquérir et les manipuler suppose donc des capacités cognitives

générales très sophistiquées. La maîtrise de telles associations "signifiant/signifié" semble accessible à certaines espèces animales (principalement les singes) à qui on a pu enseigner l'usage d'un répertoire non négligeable de symboles (gestes empruntés à une langue des signes ou dessins abstraits arbitraires). Mais aucune espèce autre que l'homme n'a développé cette capacité dans la nature, sans enseignement explicite. Quant aux concepts abstraits qui n'ont pas de référents directs dans la nature, seuls les humains semblent capables d'en faire usage.

Ce premier niveau de sens n'épuise pourtant pas toute la capacité sémantique des langues naturelles. En effet, non seulement elles offrent le moyen de désigner des choses, des individus ou des événements éventuellement distants dans le temps et l'espace, mais elles permettent surtout de construire des *énoncés* les concernant. Pour traduire la sémantique des *propositions*, c'est-à-dire des énoncés auxquels on peut attribuer une *valeur de vérité*, il faut faire appel à une combinaison de sens plus complexe : celle que, depuis Frege, on désigne par la *prédication*. Il est douteux que d'autres espèces que les hommes aient la capacité d'accéder à la prédication. C'est en tout cas ce que soutient Bickerton (Bickerton, 1990).

2.1.2 Les niveaux d'analyse

Pour représenter de façon synthétique la hiérarchie des niveaux d'analyse auxquels se prêtent les langues naturelles, nous proposons le schéma de la Figure 2.1.

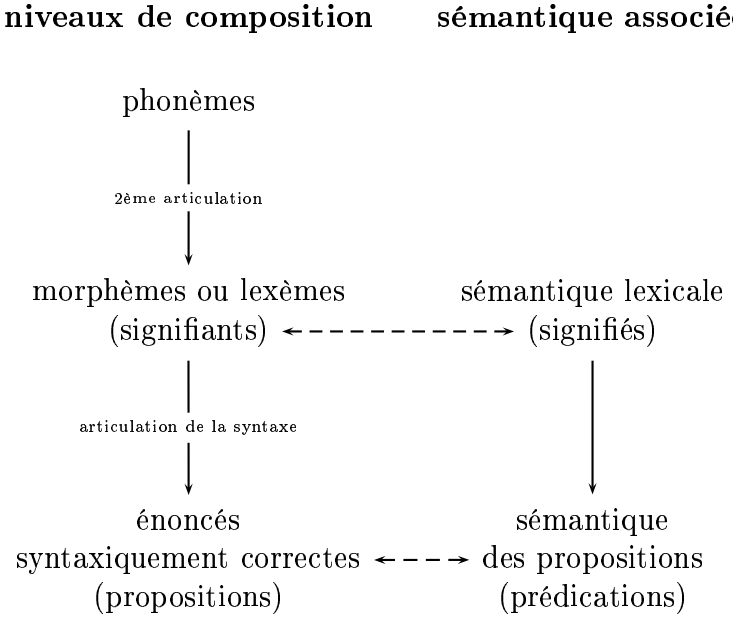

FIG. 2.1 – hiérarchie des niveaux d'analyse des langues naturelles

Dans ce schéma, les unités d'analyse figurent de haut en bas, des plus simples aux plus complexes. Les flèches verticales descendantes symbolisent donc des *règles de composition* qu'on se gardera, pour l'instant, de préciser. Les flèches bidirectionnelles en pointillés traduisent, elles, des relations d'association qui ne sont pas bi-univoques (puisqu'elles donnent lieu à des *ambiguïtés*) entre la dimension "formelle" et la dimension "sémantique" de toute langue naturelle. Un schéma plus complet devrait aussi faire figurer, à titre de niveau de composition intermédiaire, les unités de la *morphologie*. Ne sont pas évoqués ici non plus les niveaux d'analyse qui vont au-delà des propositions (analyse des discours, des textes ou des dialogues, pragmatique...). Pourtant, les hommes se racontent en permanence des histoires et des anecdotes et c'est sans doute finalement ce *comportement narratif* qui caractérise le mieux l'espèce humaine par rapport aux autres (Dessalles, 2000). Mais la Figure 2.1, et tout particulièrement le "rectangle" qu'elle fait apparaître à sa base, circonscrit en quelque sorte notre périmètre de travail.

Si ce schéma ne prétend pas épuiser tous les aspects de l'analyse du langage, il permet de bien situer les unes par rapport aux autres les distinctions évoquées précédemment, qui se focalisent chacune sur une portion de sa structure. Les modes de communication accessibles aux animaux exploitent au maximum un seul niveau de composition, celui de la "2ème articulation" de Martinet et sa sémantique associée. Mais ils sont incapables d'exprimer des énoncés propositionnels.

Les langues formelles des informaticiens (notamment les langages de programmation) peuvent, elles, être doublement articulées et on peut leur associer une "sémantique formelle". Mais la notion de sémantique envisagée alors n'a rien à voir avec celle véhiculée par les langues naturelles : son domaine est celui du calcul opérationnel, son "monde" est réduit à l'arithmétique des ordinateurs. Cette sémantique-là ne se formule pas en termes de catégories et exclut toute ambiguïté. C'est peut-être leur caractère flou et ambigu qui caractérise le mieux les langues naturelles, à ce niveau. Nous repoussons néanmoins à la partie 3, qui traitera des limites de la démarche de modélisation du langage en général (et de modélisation du sens en particulier), la question de la validité des représentations propositionnelles issues de la logique pour traduire la notion de sémantique évoquée dans ce schéma.

L'hypothèse fondatrice de notre travail est que c'est en fait *l'ensemble de cette structure qui caractérise les langues naturelles.*

2.1.3 Le Principe de compositionnalité

Pour que l'échafaudage de la Figure 2.1 ne soit pas un assemblage branlant, il faut s'assurer de la solidité des "articulations" de sa structure. C'est, pour nous, le rôle fondamental que joue le "Principe de compositionnalité". Ce principe a longtemps été attribué à Frege -en partie à tort, si l'on en croit (Janssen, 1997). Dans

une formulation contemporaine, il stipule que "le sens d'une expression composée ne dépend que du sens de ses composants et des règles par lesquelles ils sont composés" (Partee, 1990). C'est un ressort très puissant -et souvent laissé implicite- de l'analyse linguistique. Nous verrons en section 3.3 les traductions formelles auxquelles il a donné lieu, principalement suite aux travaux de Montague (Montague, 1974).

Le schéma de la Figure 2.1 permet déjà d'apercevoir le rôle qu'il peut jouer. Une proposition est, bien sûr, une "expression composée" particulière dont les "composants" sont des morphèmes. Le Principe de compositionnalité appliqué à ce niveau implique donc que le sens d'une proposition ne dépend que du sens des morphèmes qui la composent et de sa structure syntaxique, autrement dit que le "coin inférieur droit" de notre rectangle est univoquement déterminé par la connaissance des trois autres "coins". Cela signifie aussi que les flèches qui constituent les côtés opposés de ce rectangle dépendent les unes des autres.

De nombreuses critiques ont été émises à l'encontre du Principe de compositionnalité. Par exemple, il ne permet apparemment pas de rendre compte du sens des expressions figées ou idiomatiques. Mais c'est que de telles expressions doivent être considérées comme des "composants" indivisibles auxquels il faut attribuer un sens atomique. Le numéro spécial de la revue TAL qui a été consacré au Principe de compositionnalité en 1998 (Nazarenko, 1998) recense des arguments plus sérieux. Françoise Gayral et d'autres auteurs y énumèrent plusieurs problèmes : tout d'abord, le Principe de compositionnalité impose une notion de "sens lexical" qui ne laisse pas de place à la "créativité du sens en contexte" (emplois métaphoriques ou métonymiques, par exemple), à moins de formaliser *a priori* tous les emplois possibles de tous les mots. Effectivement, les modélisations du sens lexical auxquelles nous nous restreindrons se contentent d'une notion implicite de "sens propre" pour chaque mot, qui est contestable. De même, il oblige à ce que, pour déterminer le sens d'un constituant linguistique quelconque, on ne puisse prendre en compte que ses sous-constituants, alors que des effets de contexte entre constituants "frères" ou "cousins" dans une structure hiérarchique (par exemple entre un verbe et ses arguments) sont possibles. Enfin, il interdit que le sens d'un énoncé puisse dépendre de son contexte extra-linguistique (connaissances pragmatiques). Pourtant, l'interprétation des anaphores et des déictiques nécessite la prise en compte de données contextuelles non directement présentes dans les textes (qui parle, à quel moment, en s'adressant à qui...).

Tous ces problèmes, qu'ils s'en prennent au niveau des unités lexicales, des constituants ou des énoncés, sont radicalisés dans une formule attribuée à François Rastier : "c'est le global qui détermine le local, et non l'inverse". Il semble pourtant que Montague les avait largement anticipés, comme le montre Janssen (Janssen, 1997). Ils sont d'ailleurs la raison d'être principale de la notion de sens

"intensionnel" qu'il a bâti. Dans son modèle, le sens intensionnel d'une unité lexicale, d'un constituant ou d'un énoncé est en effet paramétré par des indices de monde possible et de temporalité qui permettent de recenser tous ses usages possibles. Par souci de simplification, nous avons renoncé dans nos articles à cette notion de sens intensionnel, mais elle pourrait assez facilement y être réintégrée.

Malgré ces limites, le Principe de compositionnalité est d'une grande fécondité. Il permet notamment d'expliquer comment tout locuteur d'une langue parvient à comprendre, c'est-à-dire à associer un sens, à une phrase de cette langue qu'il n'a jamais entendue ou lue auparavant. Certes, il est aussi, souvent, possible d'interpréter le sens d'un énoncé qui n'est pas parfaitement grammaticalement correct. Mais même si cette constatation va dans le sens de la primauté de la sémantique, nous ne renonçons pas pour autant à la notion de syntaxe. Le Principe de compositionnalité est aussi sûrement pour quelque chose dans l'explosion du nombre de phrases produites par un enfant vers 3 ans (cf. partie 2.2.4). Le mettre au centre de l'analyse des langues naturelles est une alternative séduisante à la thèse chomskienne de la primauté de la syntaxe car il nous semble prendre nettement mieux en compte que cette dernière toutes les dimensions caractéristiques de ces langues.

Remarquons pour finir que le Principe de compositionnalité peut également être opérant pour d'autres niveaux d'analyse : en morphologie constructionnelle, il est ainsi courant de supposer que le sens d'un mot construit par *affixation* ne dépend que du sens des morphèmes qui le composent et de leur combinaison. Cela explique que l'on peut "deviner" le sens de mots inconnus construits de façon régulière (comme "anticonstitutionnellement"). Mais la formalisation est moins avancée dans ce domaine, et nous ne l'aborderons pas plus avant. De même, certains travaux ont cherché à étendre la compositionnalité au niveau du discours et non plus à celui du seul énoncé propositionnel (Muskens, 1994), mais nous ne les détaillerons pas non plus.

2.2 Aux origines des langues et du langage

Comment un mécanisme aussi complexe que le langage, avec tous ses niveaux d'imbrication tels que décrits par le schéma de la Figure 2.1, a-t-il pu émerger dans une espèce et une seule au cours de l'évolution ? Et comment les enfants parviennent-ils à acquérir, en quelques années à peine, ce dispositif ? Comment introduire de la chronologie dans notre schéma ? Dans ce domaine comme dans d'autres, les énigmes de l'ontogénie rejoignent, en partie au moins, ceux de la phylogénie. Ce sera l'occasion de commencer à voir quel rôle l'hypothèse de la primauté du sens va pouvoir jouer.

2.2.1 Emergence

Les XVIII et XIXème siècles ont été riches en théories sur l'origine du langage. Le caractère très spéculatif -et souvent fantaisiste- de ces premières tentatives d'explication a discrédité le thème comme relevant de la recherche scientifique, au point qu'en 1866 la Société Linguistique de Paris a interdit la publication de mémoires sur le sujet. Ce n'est que depuis les années 1980 que le tabou est levé et que des conférences sur l'origine des langues et du langage se tiennent régulièrement, réunissant paléontologues, primatologues et spécialistes des sciences cognitives (y compris informaticiens).

Une des hypothèses les plus originales issue des travaux récents postule une forme intermédiaire ayant servi de jalon dans l'émergence du langage, appelée *protolangage* (Bickerton, 1990). Le protolangage est un "langage sans syntaxe", qui autorise simplement une juxtaposition non structurée de signes. Il est encore présent à l'état de fossile dans certaines formes de communication humaine : c'est le "langage de Tarzan", le parler "petit nègre" que des hommes de langues différentes emploient spontanément quand ils sont mis en situation d'échange (commerciaux par exemple) -dans ce cas, les linguistes parlent plutôt de *pidgin* ou de *sabir*. Il correspond à peu près aux compétences linguistiques d'un enfant de 2 ans, et constitue ce qu'on peut espérer de mieux d'enfants "sauvages" au sens large, c'est-à-dire ayant grandi en dehors de toute interaction langagière jusqu'à l'âge de 6 ans environ (Pinker, 1995). Comme on l'a déjà signalé, il semble accessible à certains singes. Bickerton en dote l'homo erectus (-1.6 Ma à -290 000 ans) mais pas l'homo habilis (-2.5 à -1.6 Ma). Pour l'émergence de la syntaxe, il faudrait en revanche attendre homo sapiens (apparu entre -290 000 et -140 000 ans).

Clairement, la notion de protolangage décrite ci-dessus correspond au premier niveau de notre schéma, celui où les morphèmes sont mis en correspondance avec la sémantique lexicale. A la juxtaposition de morphèmes, on peut faire correspondre une juxtaposition d'unités sémantiques, suffisante pour exprimer des combinaisons de sens élémentaires. Des simulations informatiques ont été proposées pour reproduire l'émergence de conventions linguistiques de ce niveau à l'intérieur d'une population d'agents (Kaplan, 2001; Popescu-Belis, 1999).

Dans son livre, Bickerton énumère ce qui, à ses yeux, manque au protolangage et caractérise la syntaxe des langues : l'ordre des mots, la présence d'éléments inexprimés (pronoms nuls, traces), le nombre fixe d'arguments pour chaque verbe, la construction récursive des syntagmes et l'omniprésence d'items grammaticaux. Il postule qu'une mutation unique, de type "catastrophique", peut être à l'origine de l'apparition de l'ensemble de ces traits, mais il ne propose aucun mécanisme précis pour en rendre compte. Il insiste aussi pour faire de la faculté de *représentation de la réalité* l'origine des capacités linguistiques, ce qui est une autre manière de désigner la primauté du sens.

Nous retenons de l'hypothèse du protolangage la leçon suivante : on peut postuler une étape intermédiaire dans l'émergence du langage, qui précède l'apparition de la syntaxe.

2.2.2 Les conditions de l'apprentissage d'une langue chez les enfants

L'acquisition par les enfants de leur langue maternelle est un processus universel, très étudié mais encore largement mystérieux, qui a dans les années 70 servi de catalyseur à des travaux novateurs (Piatelli-Palmarini, 1979). Comme dans beaucoup d'autres domaines liés au développement comportemental, la question de la part d'inné et d'acquis dans ce processus est un sujet très polémique. Avant d'en arriver à ce débat, nous essayons de résumer ici ce qui est connu des conditions qui rendent possibles cette acquisition.

L'histoire rapporte plusieurs récits ou légendes de rois qui, voulant connaître la langue originelle de l'humanité, auraient isolé des enfants de toute interaction langagière afin d'observer laquelle ils parlent "spontanément". Il est bien sûr impossible, pour des raisons éthiques, de reproduire de telles expériences pour explorer les "conditions limites" dans lesquelles l'apprentissage du langage est possible. Mais l'observation de diverses situations malheureuses (enfants sauvages ou enfermés dans des placards) tend néanmoins à montrer qu'un individu qui n'a jamais été exposé à aucune langue avant l'âge limite de 6 ans ne dépassera jamais, dans son développement ultérieur, le stade du "protolangage" évoqué plus haut (Pinker, 1995). Cet âge constitue donc la borne d'une *période critique* au cours de laquelle se mettent en place les fondements de l'acquisition de la langue.

Diverses autres conditions semblent *nécessaires* à l'apprentissage, en particulier l'accès au *sens* de ce qui est dit et les *interactions dialoguées* avec l'environnement (on n'apprend pas une langue étrangère simplement en regardant la télévision) (Pinker, 1994). Piaget, déjà, insistait sur le caractère actif de tout apprentissage (Piatelli-Palmarini, 1979).

Au chapitre des conditions *suffisantes*, on notera que l'acquisition est possible en présence d'*exemples positifs seuls*, c'est-à-dire de phrases syntaxiquement correctes, à l'exclusion de toute autre -ce qui veut dire, aussi, sans feed-back négatif en cas de mauvaise production de l'enfant (Wexler & Culicover, 1980). Les parents, bien sûr, peuvent pratiquer ce feed-back mais dans les civilisations où il se pratique peu ou pas du tout, les enfants acquièrent leur langue maternelle à la même vitesse que dans les autres. Ce point, souvent discuté, est particulièrement important, parce que l'apprentissage par exemples positifs seuls est notoirement beaucoup plus difficile que l'apprentissage par exemples positifs et négatifs (nous y reviendrons dans la partie 4). On peut néanmoins argumenter que certaines si-

tuations pragmatiques, mises en scène dans les dialogues où l'enfant est impliqué, tiennent lieu de feed-back négatif : quand, par exemple, il fait une demande qui n'est pas comprise. Il est de toute façon remarquable qu'en n'ayant principalement accès qu'au résultat d'une série de compositions complexes (c'est-à-dire aux données se situant au niveau le plus bas de notre schéma) et sans leçon explicite, un enfant parvienne à reconstituer le dispositif qui les a produits.

Enfin, certaines circonstances fréquentes comme, dans les civilisations occidentales, le "parler bébé" que les mères emploient à destination des jeunes enfants (les anglo-saxons parlent du "motherese"), en exagérant les contrastes acoustiques de leurs productions, constituent plutôt des *conditions contingentes* qui ne sont pas indispensables à l'apprentissage. D'autres auteurs insistent sur l'importance de facteurs affectifs (Cyrulnik, 1995) ou sociaux (Florin, 1999), mais nous n'aborderons pas plus avant ces aspects par la suite.

2.2.3 Le débat inné/acquis

Comment démêler, dans ces conditions, ce qui relève de dispositions innées de ce qui relève de comportements acquis ? Le fait que chaque langue soit différente et bâtie sur de nombreuses conventions arbitraires semble plaider en faveur de son caractère acquis. Mais c'est oublier les *universaux* qui traversent toutes les langues.

Le lien entre ces universaux et l'information génétique transmise dans le passage des générations n'est évidemment pas facile à établir. Pourtant, un premier pas a été franchi dans ce sens avec la découverte récente d'un gène (dénommé FOXP2) dont une mutation semble responsable de difficultés langagières (Lai et al., 2001; Enard et al., 2002). Et le camp des "innéistes", plutôt dominant dans les travaux actuels, dispose de nombreux autres arguments. Nous avons déjà évoqué certains d'entre eux, comme le caractère universel et spécifique à l'espèce humaine du langage et la découverte d'une "période critique" qui conditionne le bon déroulement de l'apprentissage. On peut ajouter que la chronologie des acquisitions (que nous détaillerons dans la partie suivante) est très stable, et concerne toutes les langues (y compris les langues des signes).

Historiquement, l'argument qui a été le plus discuté est celui de la "pauvreté du stimulus", qui insiste sur la faible quantité des informations disponibles à l'enfant pour lui permettre d'apprendre sa langue (Chomsky, 2001; Christophe, 2002). Par exemple, certaines constructions syntaxiques (enchassement de relatives) sont relativement peu présentes dans le langage des parents. Et les mauvaises généralisations que l'enfant pourrait effectuer ne sont pas démenties par des exemples négatifs. S'il parvient malgré tout à parler la langue de sa communauté, argumentent ces auteurs, c'est que les informations manquantes sont compensées par un dispositif inné.

Récemment, deux nouvelles séries d'observations sont venu étayer la thèse innéiste. La première concerne la transformation des pidgins en créoles. Les pidgins, on l'a déjà évoqué, sont ces pseudo-langues qui émergent quand des populations adultes de langues différentes entrent en contact. Lorsque le contact se prolonge et que des enfants naissent dans cet environnement (comme cela a été le cas lors des transferts d'esclaves aux antilles), le code a tendance à se régulariser : les pidgins évoluent et deviennent de vraies langues, des langues créoles. Selon Bickerton, spécialiste reconnu du domaine, les enfants sont les agents de cette évolution qui peut se produire en une seule génération.

L'autre série d'observations a pour cadre le Nicaragua dans les années 80. C'est un des très rares cas où on a pu observer quasiment "en direct" la naissance d'une nouvelle langue. C'est une langue des signes, et elle est née dans une cour de récréation où se trouvaient réunis des enfants sourds auxquels aucune autre langue des signes n'était enseignée. Ici aussi, quelques années ont suffi et plus les enfants étaient mis en contact jeunes, plus leur production était régulière et syntaxiquement élaborée.

Toutes ces observations (et d'autres que nous ne détaillerons pas) (Pinker, 1994) tendent donc à prouver que l'être humain dispose d'une capacité, non seulement à apprendre, mais à *créer une langue*, capacité qui s'exprime naturellement quand il est dans un environnement favorable. Apprendre sa langue maternelle, finalement, ce serait comme la réinventer à chaque génération. Chomsky, qui est bien sûr le champion de la cause innéiste, postule ainsi l'existence d'un "organe du langage" (Language Acquisition Device), présent dès la naissance. Toute sa carrière linguistique peut être vue comme une tentative pour donner un contour de plus en plus précis à cet "organe". Dans les années 60, il prenait la forme d'une "Grammaire Universelle" dont chaque langue humaine n'était qu'une instance particulière. Il a depuis pris le profil de la théorie "PP" ("Principle and Parameters"), selon laquelle l'acquisition se ramènerait à fixer la valeur d'un nombre fini de paramètres, tandis que les "principes", généraux et universels, seraient, eux, innés.

Cette position nous semble un peu extrémiste, et surtout réductrice pour le peu de cas qu'elle fait de la sémantique. De plus, les diverses listes possibles de "principes" proposées ne font pas consensus, et leur lien avec un quelconque patrimoine génétique ou biologique reste à prouver. Nous préférons, avec Pinker, parler d'un "instinct du langage" (plutôt que d'un organe), c'est-à-dire d'une disposition à l'acquisition qui demande à s'instancier par confrontation avec des données de nature *aussi bien syntaxique que sémantique*. Le seul principe auquel nous accorderons une place particulière est le Principe de compositionnalité. Des simulations récentes tendent à montrer que les langues fondées sur ce principe ont de meilleures chances que les autres d'être sélectionnées par une population d'agents artificiels qui a des besoins communicationnels (Kirby, 2002).

Nous retenons aussi des arguments de Chomsky, que l'acquisition d'une langue est plutôt un *processus de spécialisation*, qui vise à instancier des universaux dans chaque langue humaine particulière. Cette proposition va dans le sens des théories les plus récentes en psychologie du comportement, selon lesquelles l'apprentissage peut être vu comme une *inhibition progessive* de certaines possibilités présentes à la naissance (Mehler & Dupoux, 1995; Houdé, 1998).

2.2.4 Chronologie des acquisitions

Les enfants acquièrent leur langue maternelle suivant une progression homogène qui dépend très peu des individus, de la culture dans laquelle ils vivent et de la langue elle-même. La chronologie de cette acquisition est résumée dans la Figure 2.2 reprise de (Popescu-Belis, 1999), lui-même inspiré de (Kandel et al., 1995). Ces auteurs précisent que la division en capacités phonétiques, lexicales et syntaxiques est parfois difficile à établir, et qu'il y a souvent un écart temporel important entre les compétences en compréhension (plus précoces) et celles observables en production, dont ce tableau ne rend pas compte.

âge	**capacités phonétiques**	**capacités lexicales**	**capacités syntaxiques**
6-15 sem.	début du *babil*		
3-8 mois	*babil* riche		
1 an	le *babil* s'estompe ; qq. exclamations	4-5 *fonctions* pour les exclamations	
1 an 1/2	pauvreté (contrastant avec le *babil*)	30 à 50 mots : noms, adjectifs, verbes d'action	*holophrases* (phrases à un mot)
2 ans	lente amélioration : état provisoire	50 à quelques centaines de mots	*stlyle télégraphique* (phrases à 2 mots)
2 ans 1/2	idem	700 à 800 mots (proportion de noms 4 fois supérieure à celle de l'adulte)	phrases à 3 mots et plus ; nombreuses fautes
3 ans	presque adulte	un millier et plus	phrases bien formées
4 ans	quasi adulte	proche de l'adulte : env. 3000 mots (adulte : 10000 mots)	proche de l'adulte

FIG. 2.2 – chronologie de l'acquisition du langage, d'après (Kandel et al., 1995)

Il n'est pas facile de distinguer des étapes claires dans cette évolution apparemment continue (de Boysson-Bardie, 1999). Certains auteurs (Hirsh-Pasek et al., 1995; Gernsbacher, 1990), synthétisés dans (Dikovsky, 2004) (et très simplifiée ici) proposent pourtant une périodisation en 3 phases, qui prend aussi en compte le dévelopement des capacités cognitives des enfants :

- Entre 0 et 9 mois, ce sont surtout les aspects *acoustiques* de la langue que l'enfant apprend à identifier, et qu'il expérimente via son *babil*. Les données linguistiques auxquelles il est soumis semblent l'aider aussi à *segmenter* son environnement.
- Entre 9 mois et 2 ans, l'enfant apprend à mettre en correspondance les mots avec leur référent, et commence à constituer son répertoire de *catégories syntaxiques*.
- Entre 2 ans et 3 ans 1/2, il acquiert des rudiments de syntaxe, et sa représentation interne du monde s'enrichit.

Le stade "deux mots" ou "télégraphique" (vers 2 ans) est un état intermédiaire remarquable. Le langage enfantin de cette période est celui qui se rapproche le plus de l'état de *protolangage* dans le domaine de l'évolution des langues. Les productions à deux mots sont soit des combinaisons de mots lexicaux, soit une combinaison entre un mot lexical et un mot grammatical. Rappelons que les *mots lexicaux* appartiennent aux classes ouvertes du langage (noms communs, verbes, adjectifs...) et qu'on peut leur trouver un référent plus ou moins direct dans le monde (chose, personne, propriété, action, état ou événement). Les autres mots, dits aussi *grammaticaux* (prépositions, conjonctions, pronoms...) appartiennent, eux, à des classes fermées (on ne peut en inventer de nouveaux) et leur sens est, en général, moins référentiel. Les mots lexicaux sont appris en premier mais les autres apparaissent aussi assez précocement.

Si le *volume de lexique acquis* explose entre 1 an 1/2 et 4 ans (en moyenne, 3 mots nouveaux par jour), la *longueur moyenne des productions*, elle, croît en général par paliers. Dikovsky estime que la *structure sémantique du langage* est acquise entre 2 ans 1/2 et 3 ans 1/2, tandis que la *maîtrise de la syntaxe* intervient vers 3-4 ans. Nous souscrivons avec lui à la thèse du *semantic bootstraping*, initialement due à Grimshaw et Pinker (Grimshaw, 1981; Pinker, 1984), selon laquelle les connaissances sémantiques aident (et même sont indispensables) à l'acquisition de la syntaxe. Le pivot de cette aide sera le Principe de compositionnalité, comme on le détaillera par la suite.

Le processus d'acquisition que nous souhaitons modéliser est donc celui qui prend effet à partir de 2 ans environ, quand la sémantique lexicale commence à être en place mais que la syntaxe est encore balbutiante. Notons que les fautes syntaxiques commises avant 4 ans semblent souvent dues à un effet de *surgénéralisation* (par exemple conjugaison de verbes irréguliers comme s'ils étaient réguliers).

La surgénéralisation est un risque classique lors de l'apprentissage par exemples positifs seuls.

Dikovsky propose même une "grammaire archétypale" valable pour toutes les langues naturelles et qui constitue, selon lui, les connaissances syntaxiques d'un enfant de 2 ans (Dikovsky, 2004). Son effort a produit ce qu'il y a sans doute de plus avancé à l'heure actuelle dans le domaine de la formalisation du langage enfantin, même s'il manque encore à son approche une prise en compte des mécanismes d'apprentissage. Mais nous n'adopterons pas les mêmes représentations que lui, et nous nous en tiendrons à des modèles du langage qui sont traditionnels dans le champ de l'informatique linguistique, et que nous allons introduire maintenant.

Chapitre 3

Modéliser le langage

La conception du langage que nous venons d'exposer relève principalement de la linguistique. Pour modéliser le processus d'acquisition de la syntaxe dans la perspective annoncée, nous devons maintenant appliquer la démarche du schéma de la Figure 1.1 c'est-à-dire, dans un premier temps, trouver des équivalents formels aux *données* qui interviennent dans ce processus. La section suivante sera, elle, consacrée à la modélisation de l'apprentissage en tant que tel. Or, nous avons vu que les données relèvent de différents niveaux d'analyse. Nous devons donc modéliser chacun de ces niveaux, en favorisant les modèles qui permettent leur *intégration* et la mise en évidence de leurs *intéractions*.

3.1 Modéliser la syntaxe

Pour modéliser la syntaxe, les informaticiens disposent d'un outil tout prêt : la théorie des langages formels, et la hiérarchie de Chomsky qui l'accompagne (Wolper, 1997). Mais cet outil est-il adapté à notre objectif, autrement dit ce que modélise la théorie des langages coincide-t-il avec la notion de syntaxe pertinente en linguistique ? Nous verrons dans cette partie que ce n'est peut-être pas si évident que cela, et que nous devrons nous contenter de modèles partiels et imparfaits.

3.1.1 La théorie des langages formels suffit-elle ?

Un grammaire formelle au sens de Chomsky (Chomsky, 1957), c'est un dispositif calculable capable de rendre des *jugements de grammaticalité*, autrement dit de décider en un temps fini si oui ou non la suite de mots qu'on lui soumet correspond à une phrase grammaticalement correcte d'une langue. Ce jugement est, bien sûr, indépendant du caractère sensé ou non de la phrase en question, comme l'illustre le célèbrissime exemple "d'incolores idées vertes dorment furieusement", absurde

bien que grammaticalement irréprochable. Chomsky établit aussi une distinction fondamentale entre la *compétence* (connaissance théorique des règles) et la *performance* (leur usage en contexte) des locuteurs d'une langue. C'est la compétence seule qui retient son attention, et que les grammaires formelles cherchent à modéliser, tandis que la performance est soumise aux aléas des capacités cognitives individuelles.

Une grammaire formelle G au sens traditionnel est un quintuplet $G = \langle \Sigma, N, P, S \rangle$ où Σ est le vocabulaire terminal de G (constitué, pour nous, de morphèmes ou de mots), N (qui contient l'axiome S) son vocabulaire non terminal et P l'ensemble de ses règles de réécriture. Utilisée comme un *modèle génératif*, un telle grammaire produit en fait deux sortes de sorties : d'une part un *langage de chaînes* $L(G)$, qui est un sous-ensemble de Σ^*, d'autre part des *structures d'analyse*, déterminées par les séquences d'application des règles qui engendrent les éléments de $L(G)$ (souvent des arbres mais pas toujours).

Les grammaires formelles se focalisent donc sur trois aspects de la syntaxe :

- l'affectation de *catégories syntaxiques* (représentées par les symboles non terminaux) aux mots : par définition, des mots ou des groupes de mots appartiennent à la même catégorie syntaxique s'ils sont *substituables* les uns aux autres en préservant le critère de *grammaticalité* de la phrase ;
- *l'ordre des mots* (qui n'a pourtant pas la même importance dans toutes les langues naturelles) ;
- la *structuration hiérarchique* sous-jacente à cet ordre des mots.

Pour juger de la pertinence linguistique d'une grammaire donnée, il faudra prendre en compte ces trois critères et non le simple jugement de grammaticalité final qu'ils permettent d'obtenir. La notion de "strong generative capacity", suivant laquelle ce que produit une grammaire est un ensemble de *structures syntaxiques* et pas seulement un ensemble de chaînes, nous y aidera. (Miller, 1999) montre que cette notion peut aussi être vue comme spécifiant des *modèles*, ce que nous reprendrons en section 3.2.2 au niveau sémantique.

Un logicien comme Montague a pu "nier qu'il existe une quelconque différence théorique importante entre les langues formelles et les langues naturelles" (Montague, 1974). Pourtant, les propriétés que nous venons d'évoquer n'épuisent pas complètement ce que les linguistes rattachent à la syntaxe (cf. ce qui distingue un langage d'un protolangage au sens de Bickerton, partie 2.2.1). La différence entre mots lexicaux et mots grammaticaux, si importante en sémantique des langues naturelles (cf partie 3.2), est complètement absente de la théorie des langages formels. Il est aussi, par exemple, très difficile d'exprimer dans des règles qui ne surgénèrent pas le rôle et le fonctionnement des *pronoms* (les déictiques français posent de nombreux problèmes), des *appositions* ou des *ellipses*, pourtant très employés. Enfin, les grammaires formelles qu'on vient de définir sont clairement

mieux adaptées à certaines langues qu'à d'autres : les langues à ordre fixe (comme le français ou à l'anglais) se prêtent mieux à cette représentation que les langues à déclinaisons (l'allemand ou le russe) ou agglutinantes (le turc).

De nombreuses alternatives aux grammaires "classiques" existent ; tout en étant définies de manière formelle, elles se veulent plus adaptées à la représentation des langues naturelles (Sabah, 1990; Wehrli, 1997). Parmi celles qui connaissent encore des développements à l'heure actuelle, on a les *grammaires de dépendances* (Dikovsky, 2000) qui associent à la chaîne de mots un ensemble de relations de dépendances qui ne coincide pas nécessairement avec la structure produite par les règles de réécriture. Les grammaires minimalistes (Stabler, 2001) s'attachent, elles, à décrire finement les mouvements de composants à l'intérieur des syntagmes ou d'un syntagme à un autre. Les *TAG* (Joshi & Schabes, 1997) explicitent le moyen de combiner des arbres et produisent une structure de dérivation qui est plus abstraite que la structure syntagmatique. Les grammaires définies par un ensemble de *contraintes* permettent d'assouplir la notion de grammaticalité et de ramener l'analyse syntaxique à une recherche d'optima locaux (Blache, 2001). Diverses approches ajoutent ou substituent aux règles des relations *statistiques* (Manning, 1999), etc.

Les grammaires formelles qui seront la cible de notre modèle d'apprentissage appartiennent à une famille qui connaît aussi des développements contemporains : celle des *grammaires logiques lexicalisées*, dont nous évoquons les principales propriétés en section 3.1.3. Ces formalismes ne sont pas fondamentalement différents de celui évoqué plus haut puisqu'il existe des moyens de passer des uns aux autres, mais ils sont particulièrement employés pour la modélisation des langues naturelles (Oehrle et al., 1988; Miller & Torris, 1990) et entretiennent depuis longtemps des liens privilégiés avec le Principe de compositionnalité (Montague, 1974; Dowty et al., 1981), comme nous le verrons par la suite.

3.1.2 Place des langues naturelles dans la hiérarchie de Chomsky

A quel niveau de la hiérarchie de Chomsky appartiennent les langues naturelles ? Cette question est aussi vieille que les grammaires formelles et reste l'objet de débats à l'heure actuelle. Dès son premier livre, Chomsky procurait un argument qu'il pensait décisif contre la possibilité de représenter les langues naturelles par des grammaires régulières (ou, ce qui est équivalent, par des automates à états finis) (Chomsky, 1957). Cet argument est celui de *l'enchassement potentiellement infini de relatives* comme dans : "[le chien [que l'homme [que Jean connaît] regarde] court]". Ce phénomène, qu'on peut comparer au langage formel $a^n b^n$ (où les groupes nominaux sujets jouent le rôle de a et les verbes le rôle de b) est, effecti-

vement, non régulier. Pourtant, personne ne se hasardera à enchasser plus de trois relatives dans une même phrase (l'exemple précédent est à la limite de la compréhension humaine). Ainsi, l'argument ne tient que si l'on adhère à la distinction chomskienne entre compétence et performance. Chomsky estime en effet que les bornes numériques (comme le "trois" évoqué plus haut) proviennent des limites de la mémoire humaine et affectent donc seulement l'évaluation de sa performance.

Un autre argument, fondé sur des critères plus psychologiques, est dû à Pinker (Pinker, 1994). Il consiste à remarquer que le stockage en mémoire d'un automate à états finis représentant la syntaxe d'une langue naturelle est nécessairement redondant, puisqu'il doit contenir la description à la fois des groupes nominaux jouant un rôle de sujet et de ceux jouant le rôle de complément d'objet (et ce, quel que soit l'ordre des mots privilégié par la langue en question). Pourtant, tout locuteur généralise naturellement les constructions permises pour les groupes nominaux dans l'une ou l'autre position, ce qui semble contradictoire avec l'existence de stockages différents.

On peut ajouter à ces arguments le fait que les grammaires régulières ne produisent, en guise de structures sous-jacentes, que des *peignes* (c'est-à-dire des arbres binaires particuliers qui ne se développent que suivant une seule direction), alors que les structures syntagmatiques manipulées par les linguistes sont, au minimum, des vrais arbres. Admettons donc que les grammaires régulières ont une expressivité, aussi bien au niveau des chaînes que des structures, insuffisante. Cela n'empêche pas de nombreux travaux d'ingénierie linguistique, notamment en extraction d'information, d'user d'automates finis pour représenter des constructions partielles ou locales.

La classe des grammaires algébriques, ou context-free, strictement plus expressive que celle des grammaires régulières, est le prochain candidat sur la liste. Les grammaires algébriques produisent des structures d'analyse arborescentes, ce qui les rend apparemment plus adaptées au langage naturel. Pourtant, l'adéquation de ces grammaires à la modélisation des langues naturelles n'est pas encore parfaite. Certaines constructions du français nous mettent sur la piste des problèmes : Desclés (Desclés, 1982) évoque les constructions avec "respectivement". Les phrases comme "Jean, Paul et Marie achètent respectivement un pull, un pantalon et un blouson" (avec un nombre non borné de protagonistes) peuvent, en fait, être produites par une grammaire algébrique puisqu'elles sont de la forme $a^n c b^n$. Le problème est que toutes les grammaires algébriques qui produisent un tel langage le font avec des structures sous-jacentes linguistiquement inadéquates, parce qu'elles associent le premier protagoniste (ici "Jean") avec le dernier vêtement ("blouson") etc., alors que sémantiquement c'est avec le premier vêtement qu'il devrait être mis en relation. C'est donc bien un problème de *structuration hiérarchique* et non d'expressivité au niveau des chaînes que pointe cet exemple. Plus précisément encore,

comme l'argument fait appel à la sémantique, c'est l'hypothèse d'une *similarité de structure* entre syntaxe et sémantique (issue du Principe de compositionnalité) qui rend la construction incorrecte. Dans le même esprit, l'extraction de pronoms ("l'homme dont je crois que tu parles") demande des stuctures non projectives (c'est-à-dire non représentables par des arbres) (Emms, 1994).

Mais même l'expressivité au niveau des chaînes de certaines langues naturelles ne peut être générée par aucune grammaire algébrique. Schieber (Schieber, 1985) a ainsi exhibé une construction d'un dialecte suisse allémanique qui correspond au langage $a^n b^n c^n$. Depuis, on évoque aussi le génitif en géorgien ancien (Michaelis & Kracht, 1996), ou la numérotation en chinois (Radzinski, 1991).

Pour rendre compte de ces phénomènes, une nouvelle classe, plus large que la classe des grammaires algébriques mais strictement plus petite que celle des grammaires sensibles au contexte a progresivement émergé : celle des grammaires dites *légèrement sensibles au contexte* (mildly context-sensitive). Plusieurs formalismes d'expressivité équivalente ont été définis pour la caractériser (Michaelis, 2001). Une autre classe qui étend les langages algébriques sans se confondre pour autant avec la classe des langages légèrement sensibles au contexte, est décrite dans (Béchet et al., 2005).

Malheureusement, les grammaires avec lesquelles nous avons travaillé, et que nous présentons en partie suivante, ne produisent que les langages de chaînes algébriques (Bar Hillel et al., 1960). C'est aussi le cas des grammaires de Lambek (Pentus, 1993) et des grammaires de prégroupes (Buszkowski, 2001). Elles sont donc de ce fait des approximations imparfaites des grammaires que les enfants semblent capables d'acquérir.

3.1.3 Le BA-ba des grammaires catégorielles : les grammaires AB

Notre choix s'est donc porté sur les *grammaires catégorielles*, sous diverses variantes : grammaires de type AB (Bar Hillel et al., 1960), grammaires de Lambek (Lambek, 1958)) ou des extensions récentes : grammaires de prégroupes (Lambek, 1997). Comme les articles contiennent des descriptions détaillées de ces formalismes, nous nous contenterons de présenter rapidement ici, et de façon autant intuitive que formelle, les grammaires catégorielles les plus simples, dîtes AB en hommage à leurs concepteurs Adjukiewicz et Bar-Hillel.

Soit $\mathcal{B}$ un ensemble au plus dénombrable de *catégories de base*, parmi lesquelles figure une *catégorie distinguée* $S \in \mathcal{B}$, appelée l'axiome. L'ensemble des *catégories fondées sur* $\mathcal{B}$, noté $Cat(\mathcal{B})$, est le plus petit ensemble tel que $\mathcal{B} \subset Cat(\mathcal{B})$ et pour tout $A, B \in Cat(\mathcal{B})$ on a : $A/B \in Cat(\mathcal{B})$ et $B\backslash A \in Cat(\mathcal{B})$. Les éléments de $Cat(\mathcal{B})$ jouent pour les grammaires catégorielles le rôle des symboles non terminaux

pour les grammaires formelles traditionnelles. Mais la différence, essentielle, est que les éléments de $Cat(\mathcal{B})$ sont *structurés* : ils se présentent en effet sous la forme de *termes* au sens algébrique, construits à partir des connecteurs binaires / et \ et des éléments de $\mathcal{B}$, d'arité nulle. Dans certains articles où il est crucial de considérer les catégories de $Cat(\mathcal{B})$ comme des termes, nous adoptons pour elles une notation qui met mieux en évidence cette nature, en utilisant les connecteurs comme *préfixes* et en écrivant les catégories dans un ordre signifiant : le premier argument du connecteur est la catégorie jouant le rôle de *dénominateur* (celle qui figurait sous la fraction initiale), la deuxième est son *numérateur*. Ainsi, A/B se note alors $/(B, A)$ et $B\backslash A$ se note $\backslash(B, A)$.

Pour tout vocabulaire terminal fini Σ (dont les membres seront appelés des mots) et pour tout ensemble $\mathcal{B}$ de catégories de base ($S \in \mathcal{B}$), une *grammaire catégorielle* G est une relation finie sur $\Sigma \times Cat(\mathcal{B})$. On note $\langle v, A\rangle \in G$ l'affectation de la catégorie $A \in Cat(\mathcal{B})$ au mot $v \in \Sigma$.

Une *grammaire catégorielle de type* AB (ou simplement une GC par la suite) est une grammaire catégorielle dont les règles de réécriture sont réduites aux deux *schémas applicatifs* suivants : $\forall A, B \in Cat(\mathcal{B})$

- FA (Forward Application) : $A/B \;\; B \rightarrow A$
- BA (Backward Application) : $B \;\; B\backslash A \rightarrow A$

Ces schémas expliquent et justifient la notation des connecteurs sous forme de "fractions orientées", dans le mesure où les schémas FA et BA peuvent eux-mêmes être considérés comme des "réductions de fraction". Les noms de "Forward Application" et "Backward Application" mettent aussi en évidence le fait qu'une catégorie de la forme A/B (resp. $B\backslash A$) est une sorte de foncteur qui attend comme argument sur sa droite (resp. sur sa gauche) une catégorie B et produit comme résultat la catégorie A. La notation sous forme de terme présente l'avantage de rendre plus facilement lisible ces rôles, fixés par l'ordre des arguments des connecteurs / et \.

Le langage $L(G)$ reconnu (ou engendré) par une GC G est l'ensemble des chaînes de mots pour lesquelles il existe une affectation de catégories qui, via les schémas FA et BA, se réduit à la catégorie axiome. Formellement, $L(G)$ est donc défini par : $L(G)=\{w = v_1 \dots v_n \in \Sigma^+ \mid \forall i \in \{1, \dots, n\}, \exists A_i \in Cat(\mathcal{B})$ tel que $\langle v_i, A_i\rangle \in G$ et $A_1 \;\dots\; A_n \rightarrow^* S\}$, où $\rightarrow^*$ est la clôture réflexive et transitive de la relation définie par FA et BA. Les arbres d'analyse syntaxique correspondant sont toujours des arbres binaires dont les noeuds internes sont étiquetés par FA ou BA, en plus des étiquettes de catégories (cf. Figure 3.1).

La particularité principale des GCs est qu'elles sont *lexicalisées*, au sens où la totalité de l'information syntaxique est associée via les catégories aux mots du vocabulaire. Ce sont par exemple les connecteurs qui fixent l'ordre des mots valide dans une grammaire donnée. Les règles sont, elles, définies une fois pour toutes et invariables d'une grammaire à une autre.

Exemple 1 (Une GC très simple). *Illustrons la construction d'une GC élémentaire G pour reconnaître quelques phrases du français. Soit $\Sigma = \{Jean, dort, un, chat\}$. Pour que "Jean dort" appartienne au langage reconnu par notre grammaire, deux choix sont possibles suivant que la catégorie associée au premier mot joue le rôle de foncteur ou celui d'argument. Le premier choix consiste ainsi à affecter une catégorie de la forme S/B à "Jean" et B à "dort", tandis que l'autre choix affecte une catégorie de la forme B à "Jean" et $B\backslash S$ à "dort". Nous verrons dans la partie 3.3 que pour des raisons purement sémantiques, c'est le deuxième choix qui est préférable. Soit donc $\mathcal{B} = \{S, T, NC\}$ (où T désigne la catégorie des "termes" et NC celle des "noms communs") et les affectations suivantes : $\langle Jean, T\rangle \in G$, $\langle dort, T\backslash S\rangle \in G$ et $\langle chat, NC\rangle \in G$.*

De même, pour que la phrase "un chat dort" soit grammaticale tout en reconnaissant "un chat" comme un constituant (correspondant donc à un sous-arbre de l'arbre d'analyse de la phrase), plusieurs choix d'affectations sont possibles pour "un", comme le montre la Figure 3.1. Nous reviendrons plus tard sur les conséquences des différents choix possibles.

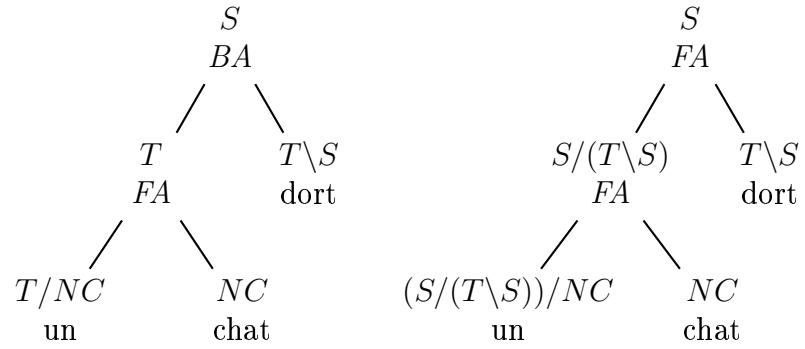

FIG. 3.1 – deux analyses syntaxiques possibles pour "un chat dort"

Les grammaires catégorielles de Lambek (Lambek, 1958) sont une extension naturelle des grammaires AB. Elles s'en distinguent par les schémas de règles qui prennent la forme *d'axiomes et de règles d'inférence*. Réaliser une analyse syntaxique est ainsi identifié à prouver un théorème logique. Dans les grammaires de prégroupe (Lambek, 1997; Lambek, 2001; Lambek & Casadio, 2002), enfin, les catégories de base sont liées les unes aux autres par une relation d'ordre partielle, compatible avec la concaténation, et chaque catégorie est associée à un pseudo-inverse à gauche et un pseudo-inverse à droite. Faire une analyse syntaxique, cette fois, revient à prouver une inégalité suivant la relation d'ordre en tenant compte de de celles déjà connues entre catégories de base, et de celles qui relient les catégories et leurs pseudo-inverses.

La famille des grammaires logiques lexicalisées compte de nombreux autres rejetons, qui s'appellent : Combinatorial Categorial Grammars (Steedman, 1996; Steedman, 2000), logique linéaire (Rétoré, 2002), grammaires multimodales (Moortgat, 1997), Abstract Categorial Grammars (de Groote, 2001; de Groote & Pogodalla, 2004). Mais n'ayant pu travailler sur l'apprenabilité de telles grammaires, nous ne les développerons pas plus avant.

3.2 Modéliser la sémantique

Jacques Arsac estimait que les modèles informatiques butaient -et allaient toujours buter- sur ce qu'il appelait le "mur du sens" (Arsac, 1987; Arsac & Vauthier, 1989). Le fait que le sens ne se réduise pas à une manipulation de symboles est aussi le fond de la pensée de Searle, dans son fameux argument de la "chambre chinoise" (Searle, 1980). Certes, sans capacités perceptives, l'ordinateur est une machine autiste; il n'a aucune connaissance du monde extérieur autre que celle qu'on lui fournit de manière explicite. Mais les connaissances humaines sur le monde sont bien, pourtant, "stockées" quelque part dans son organisme. Le pari du cognitivisme, c'est de pouvoir coder sous forme de données symboliques ces connaissances, et ainsi de les transférer sur support digital. Voyons donc comment les informaticiens ont tenté de "coder" ce que le langage dit du monde, pour le transmettre aux ordinateurs (Sabah, 1990; Rastier et al., 1997).

3.2.1 Approches de la sémantique lexicale

Comme le montre le schéma de la Figure 2.1, la modélisation du sens peut s'opérer à deux niveaux différents. Nous nous concentrons dans un premier temps sur la sémantique lexicale, et plus particulièrement sur la classe ouverte des mots lexicaux (cf. partie 2.2.4) auxquels on peut le plus facilement trouver un référent dans le monde. Pour rendre compte du sens de ces mots, on peut dire que, très schématiquement, deux principales approches ont été proposées (en dehors de la logique, qui sera développée dans la section suivante).

La première s'attache à définir des *unités de sens*, des *primitives* dont la granularité est plus petite que celle portée par une unité lexicale, et donc à associer à chaque "mot"[1] une *combinaison* de telles unités. C'est la base de la *sémantique componentielle*, qui a donné naissance à *l'analyse sémique* (Nyckees, 1998). Les unités de sens considérées dans ce cas sont des *sèmes* (du type animé/inanimé, masculin/féminin, etc.). Mais, comme il est quasiment impossible de recenser l'ensemble des sèmes qui permettraient de caractériser le lexique d'une langue complète, cette

[1] la notion de "mot" n'est pas pertinente en linguistique, mais pour des raisons de commodité, on s'autorisera néanmoins à l'employer

théorie a donné lieu a peu d'applications informatiques (à l'exception de (Rastier et al., 1997; Beust, 1998)). On peut dire pourtant que les *traits sémantiques* utilisés dans certaines grammaires d'unification sont les héritiers directs de ces sèmes. On les réduit, dans ce cas, à jouer le rôle de contraintes sur l'application de règles syntaxiques.

Plus récemment, est apparue une autre manière d'envisager le sens des mots lexicaux : la théorie du lexique génératif (Pustejovsky, 1995). Un de ses apports originaux réside dans la façon de structurer les différents aspects du sens en 4 "qualia" : le qualia "constitutif" explicite la relation entre l'objet référé et ses composantes (matière, poids, partie), le qualia "formel" caractérise ses propriétés générales (orientation, taille, forme, etc.), le qualia "télic" sa fonction et le qualia "agentif" les facteurs impliqués dans sa création. Certains de ces "rapports qualia" expliquent l'association entre un prédicat et un argument (par exemple un verbe et son sujet) et permettent, par exemple, l'extension de requêtes dans les moteurs de recherche (Claveau & Sébillot, 2004; Claveau & Sebillot, 2004). Cette théorie est prometteuse, mais elle connaît encore quelques problèmes et la conception du lexique complet d'une langue n'y est pas encore à l'ordre du jour.

En simplifiant, on peut dire que le point commun entre les théories précédentes était de considérer le sens d'un mot lexical comme une combinaison (plus ou moins structurée) de sens "élémentaires". L'autre approche, inspirée par des travaux sur la la mémoire humaine, renonce aux unités de sens pour concevoir la sémantique lexicale comme un *système* ou un *réseau*. Les noeuds du réseau peuvent être, suivant les cas, des noms communs, des termes linguistiques ou des concepts (Gruselle, 1997), et les liens qui les relient sont typés par des relations comme "sorte de", "partie de", etc. (de nombreuses variantes existent en la matière). Ce sont ces relations possibles, en nombre fini, qui constituent la structure du réseau, et le sens d'un noeud y est conçu non comme une caractéristique propre à ce noeud (quoiqu'on puisse lui attribuer des traits, qui se propagent par héritage le long de certains liens), mais comme émergeant de ses relations avec les autres. On peut rattacher à cette famille le réseau Wordnet (http ://wordnet.princeton.edu/), très employé dans le domaine de la recherche d'information, les ontologies à visée universelle (Cyc : http ://www.opencyc.org/ , Sumo : http ://www.ontologyportal.org/) et, à une échelle plus modeste, les thesaurus de spécialité utilisés dans les logiciels documentaires.

Mais ces différentes conceptions de la sémantique lexicale, que l'on n'a fait que très rapidement survoler, nous seront peu utiles par la suite. Les principaux reproches que l'on peut leur faire sont de se focaliser sur une *partie* du lexique et de ne pas pouvoir, du moins dans leur état actuel, servir de base à une sémantique de la proposition. Etant complètement déconnectées de la syntaxe, elles sont difficilement compatibles avec la prise en compte de la compositionnalité.

3.2.2 Les représentations sémantiques propositionnelles

Quand il s'agit de représenter la *sémantique des propositions*, ce sont les modèles de type *logique*, qui intègrent au moins la logique des prédicats du premier ordre, qui servent de référence. C'est en tout cas une représentation de ce type que nous utiliserons. Mais ce choix mérite discussion. La logique, historiquement, est née pour un autre usage : celui de la *formalisation du raisonnement*. Ce n'est que dans les années 1970, avec le développement de l'intelligence artificielle "classique" (Crevier, 1999) qu'elle s'est aussi imposée pour la *représentation des connaissances*. C'est de ce dernier emploi dont nous avons besoin ici.

Les enfants, en effet, apprennent à parler bien avant d'apprendre à raisonner ; les lois de la logique classique et des valeurs de vérité leur restent longtemps impénétrables (Piaget, 2003). Pourtant, leurs capacités de catégorisation et de représentation du monde se développent très tôt (Mehler & Dupoux, 1995; Pinker, 2000; Houdé, 1998). Pour donner une telle capacité aux ordinateurs, bien des formalismes ont été proposés (Kayser, 1997). Parmi les plus connus dont l'objectif était explicitement la sémantique des langues naturelles, on peut citer (Sabah, 1990) : les représentations conceptuelles de Schank (Schank & Abelson, 1977), les réseaux sémantiques (Quillian, 1968), le langage de Jackendoff (Jackendoff, 1990), les graphes conceptuels (Sowa, 1984), la théorie "situation and attitudes" (Barwise & Perry, 1983) et, plus recemment, les "meaning structures" de Dikovsky (Dikovsky, 2003).

En fait, plusieurs de ces formalismes ont une expressivité très similaire à celle de la logique des prédicats du premier ordre, ou l'étendent en affinant son typage ou en y intégrant des primitives ou des modalités. La conception du sens propositionnel qu'elles véhiculent est celle d'une *formule bien formée* issue d'un langage formel et dont l'interprétation dépend d'un *modèle* du monde, décrit en général en termes de théorie des ensembles. Les mots s'y traduisent donc par des *primitives fonctionnelles autonomes* ou des λ-termes, comme on le détaillera dans la partie suivante. L'*extension* ou la *dénotation* de ces formules est traditionnellement fournie par une *fonction d'interprétation* qui les relie aux propriétés du modèle considéré. Les deux niveaux de sens identifiés précédemment sont ainsi parfaitement intégrés dans un cadre unique.

Signalons au passage que, tout en faisant usage d'une variante logique de la représentation des connaissances, nous n'adhérons pas nécessairement à l'identification du *sens* avec un simple calcul de *conditions de vérité*. Des approches récentes, en effet, permettent d'envisager des alternatives interprétatives qui préservent la *syntaxe* de la logique des prédicats du premier ordre, mais renouvellent la façon de considérer les formules qu'elle produit. Ainsi, en "Dynamic Predicate Logic" (Groenendijk & Stockhof, 1991), un quantificateur existentiel a principalement pour rôle d'introduire un nouveau référent auquel la suite du discours pourra faire appel. En

"Update Logic" (Groenendijk et al., 1996b; Groenendijk et al., 1996a) une formule est interprétée comme un ensemble de conditions imposées au(x) modèle(s) courant(s) du monde pour qu'il(s) soi(en)t compatible(s) avec elle. Dans ces deux théories, une formule est vue comme un *modifieur de modèles*. L'enjeu n'est pas de savoir si elle est vraie ou fausse, mais plutôt de la considérer comme porteur d'une information sur le monde dont il faut tenir compte pour mettre à jour la conception qu'on en a. De même que la compétence syntaxique ne se réduit pas à un jugement de grammaticalité, la compétence sémantique ne se réduit pas à un verdict vrai/faux. Le résultat de l'effort d'interprétation est le modèle lui-même, et non une simple valeur de vérité. Cette approche est cohérente avec la théorie des *modèles mentaux*, selon laquelle nos inférences et nos raisonnements se font sur la base de modèles plus que de règles (Johnson-Laird, 1983).

La représentation sémantique propositionnelle, ainsi conçue, est en quelque sorte le *langage de spécification* des modèles du monde que l'esprit élabore. C'est à ce titre qu'on peut la rapprocher du "langage de la pensée" postulé par Fodor (Fodor, 1975). D. Israel (Israel, 2003), lui, a pu aller jusqu'à dire que tous les travaux d'informatique linguistique fondés sur des sémantiques formelles reposent sur l'hypothèse sous-jacente que les langues naturelles sont le "langage de programmation de l'esprit humain". Quoi qu'il en soit, les informaticiens ne cherchent pas en général à simuler les modèles eux-mêmes ((Hobbs & Rosenstein, 1977) propose pourtant de comparer l'état interne de l'ordinateur à un tel modèle), et s'en tiennent aux formules qui les spécifient.

Mais, si le langage de la pensée existe, les formalismes de représentation des connaissances n'en sont sans doute encore qu'une piètre approximation. Leurs limites sont flagrantes, en effet, lorsqu'il s'agit de leur faire traduire des propriétés ou des constructions pourtant présentes dans la plupart des langues naturelles : les genres, les nombres et tous les circonstants qui expriment des relations de temps, de lieu, de cause, de conditions, etc. Pour remédier à ces problèmes, on fait généralement appel à des extensions techniques : traits, modalités, connecteurs temporels... La logique intensionnelle de Montague (Montague, 1974; Dowty et al., 1981; Delsarte & Thayse, 2001) a constitué, à son époque, une synthèse, dont notre logique sera l'héritière simplifiée.

Certaines extensions plus ambitieuses comme la DRT (Kamp & Reyle, 1993) et la SDRT (Asher, 1993) ont permis de donner une première solution à l'expression de l'accessibilité anaphorique dans un discours ainsi qu'à la représentation de liens argumentatifs entre phrases. D'autres se sont focalisées par exemple sur l'interprétation des espaces fictionnels que la langue est capable de susciter (Fauconnier, 1984). Le champ des recherches en ce domaine est encore riche et vaste. Mais tout ce qui fait le sel du langage, son pouvoir évocateur et poétique, tout ce qui empêche de s'en tenir au "sens premier" d'une expression (humour, ironie, mé-

taphores et métonymies...) constituent pour toutes ces théories un horizon encore bien lointain.

En résumé, ce n'est pas de la logique en tant que telle que nous avons besoin, mais d'un mode de *représentation des connaissances* qui ait les propriétés suivantes :

- c'est un langage formel
- dont les formules bien formées représentent des propositions capables de spécifier des modèles
- dont certaines sous-formules représentent la sémantique lexicale
- qui est lié à la syntaxe des langues naturelles de manière *compositionnelle*.

Il est inutile de détailler ici les propriétés de la logique que nous avons utilisée, parce qu'elles découlent en fait de l'exigence de compositionnalité. C'est donc en modélisant le Principe du même nom, sur laquelle repose toute la solidité de notre édifice, que nous préciserons nos choix.

3.3 Modéliser la compositionnalité

Nous avons énoncé la formulation contemporaine du Principe de compositionnalité en partie 2.1.3. Elle laisse entendre que si on s'est fixé une syntaxe et une sémantique lexicale, alors la sémantique propositionnelle s'en déduit. Nous allons en fait montrer qu'il faut sans doute voir le problème à l'envers et que la vraie question que se sont posé Montague et ses héritiers est plutôt : comment fixer une syntaxe et une sémantique lexicale pour que la sémantique propositionnelle en découle de façon compositionnelle ?

3.3.1 Introduction sur un exemple

Pour introduire cette démarche, voyons comment associer une sémantique compositionnelle à l'embryon de grammaire de l'Exemple 1, partie 3.1.3. On rappelle que pour analyser la phrase "Jean dort", deux alternatives sont possibles suivant qu'on affecte une catégorie de la forme B à "Jean" et $B\backslash S$ à "dort" ou alors une catégorie de la forme S/B à "Jean" et B à "dort". Or, d'un point de vue sémantique, cette phrase se traduit par la formule logique $dort'(jean')$ où $dort'$ est un prédicat à un argument qui traduit le sens du mot "dort" et où $jean'$ est une entité qui identifie l'individu "Jean" dans le modèle. On voit alors pourquoi la première affectation de catégories, qui reproduit la structure "foncteur/argument" exhibée par la représentation sémantique, est préférable à l'autre. Ne concluons toutefois pas trop vite.

La Figure 3.1, partie 3.1.3, montre qu'un choix du même ordre doit être opéré pour analyser la phrase "un chat dort". Mais comment, dans ce cas, traduire les

mots de cette phrase pour aboutir à sa traduction naturelle en logique du premier ordre, la formule : $\exists x[chat'(x) \wedge dort'(x)]$? Le coup de génie de Montague (Montague, 1974; Dowty et al., 1981) a consisté à introduire des expressions du λ-calcul dans la logique, pour établir ce lien. L'idée est, grossièrement, que les mots "chat" et "dort", dont l'extension sémantique est un ensemble d'individus, se traduisent tous deux par un prédicat à un argument, respectivement noté $chat'$ et $dort'$ et que le reste de la formule doit donc être porté par la traduction du mot grammatical "un". L'arbre droit de la Figure 3.1 suggère d'ailleurs qu'on peut affecter à "un" une catégorie qui jouera le rôle de foncteur pour les catégories des deux autres mots. C'est aussi ce rôle qu'on peut lui faire jouer au niveau sémantique, en le traduisant par la formule : $\lambda P \lambda Q \exists x[P(x) \wedge Q(x)]$. Le "traitement correct des quantificateurs" ("Proper Treatment of Quantification" est le titre de l'article fondateur de Montague) consiste donc à les prendre pour porteurs de la structure fonctionnelle des propositions, aussi bien au niveau syntaxique que sémantique. L'arbre syntaxique sélectionné pour la phrase "un chat dort" et son pendant sémantique sont représentés dans la Figure 3.2 (où figurent également les β-réductions qui permettent d'aboutir à la formule finale, en gras).

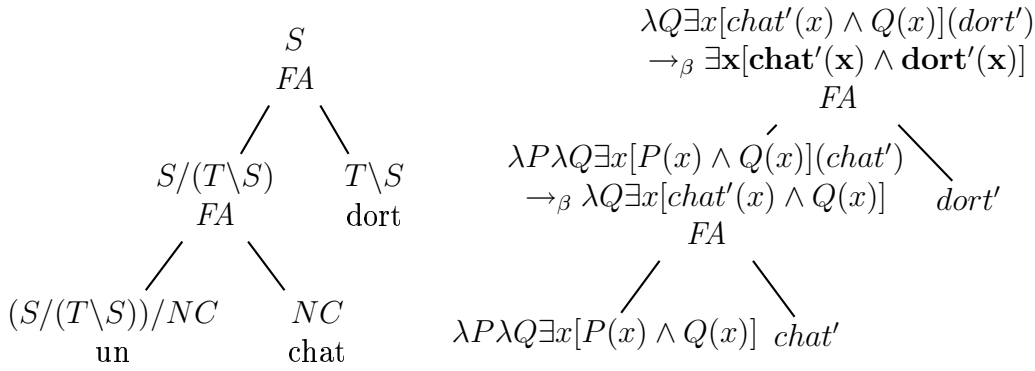

FIG. 3.2 – arbres syntaxique et sémantique de "un chat dort"

Sur la base de cet exemple, Montague, préférant un traitement homogène pour tous les groupes nominaux, décida finalement d'affecter aux "noms propres" comme "Jean" une catégorie qui leur fasse aussi jouer le rôle de foncteur, ce qui remet en cause la première conclusion que nous avions tirée. Cela n'est possible qu'à condition de les traduire par une formule qui puisse aussi se comporter comme un foncteur vis-à-vis d'un groupe verbal prédicat, d'où pour "jean" la traduction : $\lambda P(P(jean'))$ (cf. Figure 3.3). Nous estimons toutefois que ce choix n'a rien d'obligatoire, et qu'on peut très bien adopter le traitement "à la Montague" des quantificateurs sans pour autant le généraliser aux groupes nominaux comme les identifiants de personnes où, justement, aucun quantificateur n'est présent. Mais

cette digression montre surtout à quel point, dès qu'on adopte une perspective compositionnelle, les choix syntaxique et sémantique sont solidaires.

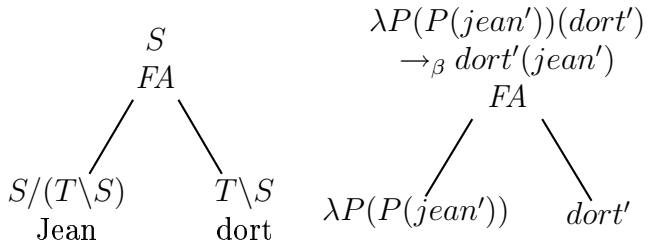

FIG. 3.3 – syntaxe et sémantique de "Jean dort" façon Montague

Notons aussi que cette perspective donne un statut sémantique aux mots grammaticaux, ce dont aucune des autres théories de la sémantique lexicale évoquées en partie 3.2.1 n'est capable. En leur attribuant une traduction à base de λ-termes, elle fait reposer sur eux la "forme logique" de la traduction d'une proposition.

Le langage de représentation des connaissances dont nous faisons usage, à la suite de Montague, est donc une logique étendue par le λ-calcul. A la suite de Montague toujours, nous adoptons aussi le *typage* qu'il a proposé pour les formules de cette logique. Ce typage, dans sa version minimale simplifiée (sans employer la notion d'intension qu'il avait introduite) repose sur deux types de base : le type e des *entités du modèle* et le type t des *valeurs de vérité*. L'ensemble de tous les types possibles, noté $\mathcal{T}$, est alors le plus petit ensemble contenant ces deux types et tel que si a et b sont des types, alors $\langle a, b\rangle$ désigne le type des formules qui prennent un argument de type a et donnent comme résultat une formule de type b. Par exemple, les prédicats à un argument comme $chat'$ et $dort'$, dont l'extension est un sous-ensemble de l'ensemble des entités, autrement dit une fonction caractéristique de cet ensemble, sont de type $\langle e, t\rangle$, tandis que les prédicats à deux arguments sont de type $\langle e, \langle e, t\rangle\rangle$. La formule $\lambda P \lambda Q \exists x[P(x) \wedge Q(x)]$ qui traduit "un", quant à elle, attend deux prédicats à une place (notés P et Q) comme arguments et produit une proposition complète ; elle est donc de type $\langle\langle e, t\rangle, \langle\langle e, t\rangle, t\rangle\rangle$.

Ce typage permet de contraindre les applications fonctionnelles possibles : un foncteur f de type $\langle a, b\rangle$ ne peut s'appliquer qu'à un argument u de type a pour donner $f(u)$, de type b. Dans une telle logique, il est impossible d'appliquer un foncteur sur lui-même : aucun typage ne permet d'écrire $f(f)$, par exemple. A l'origine, cette notion de typage était destinée à éviter le paradoxe de Russel (dans lequel un ensemble peut être élément de lui-même) sur la théorie des ensembles. Mais, on peut aussi la voir comme la *trace lexicale* du Principe de compositionnalité, ce que nous expliquons dans la section suivante.

3.3.2 Approches formelles de la compositionnalité

Dans tous les exemples précédents, on a admis implicitement que "l'application fonctionnelle" exprimée dans les schémas applicatifs *FA* et *BA*, propres aux grammaires catégorielles de type AB, avait pour exact pendant sémantique l'application fonctionnelle mathématique standard de la logique et du λ-calcul. Les analyses syntaxique et sémantique peuvent ainsi figurer dans *deux arbres isomorphes* comme l'illustrent les Figures 3.2 et 3.3. Là où le résultat syntaxique (dans les arbres à gauche) est une catégorie issue d'une réduction de fraction, le résultat sémantique (dans les arbres à droite) est la combinaison de deux sous-formules, dans l'ordre indiqué par la nature du schéma qui a été appliqué (puisque le foncteur est à gauche si *FA* s'applique, à droite sinon).

Le passage des *catégories syntaxiques* aux *types logiques* peut, alors, être vu comme un morphisme, noté h, qui met en relation deux ensembles de termes, tous les deux fondés sur une notion "d'application fonctionnelle". Aux *catégories de base* de l'ensemble $\mathcal{B}$, h commence par faire correspondre des types de l'ensemble $\mathcal{T}$, en respectant la correspondance entre phrases correctes et propositions soit : $h(S) = t$, puis il s'étend naturellement à toutes les catégories de $Cat(\mathcal{B})$ de la façon suivante (on adopte ici la notation des catégories sous forme de terme, nettement mieux adaptée) : $h(/(B, A)) = h(\backslash(B, A)) = \langle h(B), h(A)\rangle$.

Exemple 2 (traduction des catégories syntaxiques aux types logiques). *Explicitons le morphisme h sous-jacent à la relation entre la grammaire de l'Exemple 1 et les traductions de la partie 3.3.1. Pour les catégories de base, on a donc :* $h(S) = t$, $h(T) = e$ *et* $h(NC) = \langle e, t\rangle$. *Notons que rien n'interdit qu'à une catégorie de base, h associe un type qui, lui, n'est pas un type de base (comme c'est le cas pour les noms communs NC). Le type correspondant aux verbes intransitifs de catégorie* $T\backslash S$ *est alors* $h(\backslash(T, S)) = \langle h(T), h(S)\rangle = \langle e, t\rangle$. *On retrouve bien, ainsi, le fait que les noms communs comme "chat" et les verbes intransitifs comme "dort" ont le même type sémantique des prédicats à une place* $\langle e, t\rangle$, *bien qu'ils ne proviennent pas des mêmes catégories syntaxiques. Et pour le mot grammatical "un", on a bien aussi :* $h((S/(T\backslash S))/NC) = h(/(NC, /(\backslash(T, S), S))) = \langle\langle e, t\rangle, \langle\langle e, t\rangle, t\rangle\rangle$.

Un *morphisme d'arbres* fondé sur un isomorphisme de structures et une mise en correspondance des feuilles, c'est la conception de la compositionnalité héritée de Montague (Montague, 1974; Dowty et al., 1981; Janssen, 1997). A vrai dire, lui n'utilisait pas les schémas *FA* et *BA* mais d'autres règles syntaxiques *ad hoc*, auxquelles il faisait correspondre autant de règles de composition sémantique, tout en respectant le parallélisme catégories/types. Mais on peut aller plus loin. Au prix d'une réécriture des catégories de la forme A/B en $B \longrightarrow A$ et de $B\backslash A$ en $A \longleftarrow B$, les schémas *FA* et *BA* peuvent à leur tour s'écrire comme des séquents logiques :

$$[FA]\ \frac{B \longrightarrow A\ ,\ B}{A} \quad [BA]\ \frac{B\ ,\ A \longleftarrow B}{A}$$

FA et *BA* apparaissent alors comme des instances orientées du *modus ponens* de la logique classique[2] et la correspondance syntaxe/sémantique n'est rien d'autre qu'une instance de *l'isomorphisme de Curry-Howard*. Selon cette isomorphisme, toute preuve logique peut également être interprétée comme un programme, exprimé par un λ-terme (et inversement). Dans le cas qui nous intéresse, c'est surtout l'identification entre l'application fonctionnelle aux niveaux syntaxique et sémantique qui est ainsi justifiée, et qu'on peut du coup intégrer dans une "preuve" unique, dont les séquents suivants constituent les règles (où f et u sont des λ-termes "typés" par les catégories qui leur sont associées) :

$$[FA]\ \frac{f : B \longrightarrow A\ ,\ u : B}{f(u) : A} \quad [BA]\ \frac{u : B\ ,\ f : A \longleftarrow B}{f(u) : A}$$

Cette identification a été, bien sûr, développée et approfondie dans le cadre de formalismes catégoriels plus complexes comme les grammaires de Lambek (Moortgat, 1988), ses extensions multimodales (Moortgat, 1997), et la logique linéaire (Morill, 1994). Les Abstract Categorial Grammars (de Groote, 2001) sont aussi fondées sur cet isomorphisme.

Dans sa formulation standard, le Principe de compositionnalité explique comment la sémantique propositionnelle *se déduit* d'autres informations. Mais cela ne rend pas justice à une formalisation qui, elle, accorde un statut équivalent aux analyses syntaxique et sémantique, par le biais d'un *isomorphisme*. Comme notre projet est plutôt d'expliquer comment la syntaxe peut être acquise à l'aide d'informations sémantiques, nous verrons en partie 5.1 que cette approche doit s'accompagner d'une *reformulation du Principe de compositionnalité* qui "inverse" en quelque sorte son sens.

L'utilisation de la compositionnalité, modélisée par l'isomorphisme de Curry-Howard, pour déduire des informations syntaxiques à partir d'informations sémantiques est au coeur du travail de thèse de (Pogodalla, 2001). Mais ce dernier ne se situe pas dans une perspective d'apprentissage syntaxique ; il se rattache à la problématique de la *génération automatique*, où il s'agit de produire un texte (engendré par une grammaire connue) à partir d'une représentation sémantique de son sens. Les deux domaines gagneraient sans soute à être rapprochés, mais c'est une piste que nous n'avons pas eu le temps d'envisager sérieusement.

[2] c'est cette propriété qui justifie l'appellation de formalisme logique lexicalisé employé pour désigner les grammaires catégorielles en partie 3.1.3

3.3.3 Ambiguités

Reste un problème que nous n'avons pas encore abordé : celui de l'ambiguité, ou plutôt *des* ambiguités qui peuvent survenir tant au niveau de la sémantique lexicale qu'à celui de la sémantique des propositions. Comment l'affectation d'une sémantique formelle, par définition *non ambiguë*, à des énoncés de la langue naturelle qui, eux, peuvent l'être, est-elle compatible avec le Principe de compositionnalité ? C'est ce que nous allons voir.

Pour ce qui est de la sémantique lexicale, le problème n'est pas si sérieux qu'il en a l'air. D'une part, il est recommandé de définir la fonction de traduction du vocabulaire en formules logiques en prenant pour ensemble de départ $\Sigma \times Cat(\mathcal{B})$ plutôt que Σ seul. Ainsi, les mots ambigus à cause de leurs multiples assignements de catégories syntaxiques (par exemple "garde" qui peut être à la fois un nom commun et un verbe ou, pour les mots grammaticaux, "la" qui peut être un déterminant ou un pronom) peuvent être "désambiguisés" par la donnée de cette catégorie syntaxique. Pour les phrases ambiguës du fait de multiples choix de ce genre, comme le célèbre exemple chomskien "time flies like an arrow" (ou les équivalents français comme "le combattant brave la garde"), il y a *autant d'arbres syntaxiques différents qu'il y a de sens possibles.* Comme le Principe de compositionnalité opère au niveau des *structures* et non au niveau des chaînes de mots, il n'est en rien remis en cause. D'autre part, pour les mots qui ont plusieurs sens de même catégorie syntaxique (comme "avocat"), il faut évidemment considérer plusieurs traductions distinctes de même type (deux prédicats à une place respectivement notés $avocat'_1$ et $avocat'_2$).

Il existe dans les langues naturelles un autre type, subtil, d'ambiguités auquel Montague (Montague, 1974; Dowty et al., 1981) a consacré beaucoup d'efforts : on la désigne sous l'alternative *de re/de dicto.* On ne la rencontre qu'en présence de ce que, depuis Frege, on appelle des "contextes opaques", c'est-à-dire des environnements linguistiques où la substitution d'un syntagme par un autre de même extension ne préserve pas nécessairement le sens global de la phrase. Certains verbes comme "croire", "imaginer", "chercher"... introduisent de tels contextes (Galmiche, 1991; Chambreuil, 1989). Ainsi une phrase comme "Jean cherche un crayon" a deux sens possibles, suivant que Jean cherche un objet unique (un crayon précis qu'il sait avoir laissé quelque part), ou n'importe quoi pouvant tenir lieu de crayon (c'est-à-dire ayant les propriétés d'un crayon). D'après Montague, les traductions logiques respectives de ces deux interprétations, appelées "de re" et "de dicto" sont : $\exists x[crayon'(x) \wedge cherche'(jean', x)]$ et $cherche'(jean', \lambda Q \exists x[crayon'(x) \wedge Q(x)])$. Pour comprendre l'intérêt de la deuxième formule, on peut penser, comme le suggère Montague, à la phrase "Jean cherche une licorne" à laquelle on doit pouvoir attribuer la valeur de vérité "vraie" même si aucune licorne n'existe dans le monde, ce que ne permettrait pas la première formule. Pour obtenir ces fomules de ma-

nière compositionnelle, il devait introduire des règles spécifiques. Mais nous avons montré (Tellier, 1992) qu'on pouvait obtenir exactement le même résultat dans le système de Lambek, sans provoquer de surgénération, en attribuant une nouvelle catégorie syntaxique (et donc un nouveau type au prédicat correspondant) aux verbes qui introduisent des contextes opaques.

Mais les cas d'ambiguités propositionnelles les plus célèbres et les plus étudiés proviennent des problèmes de *portée des quantificateurs* (le terme anglais consacré est "scope ambiguity"). L'exemple prototypique est fourni par la phrase "chaque homme aime une femme" mais on peut lui préférer "chaque élève apprend une langue", qui a deux interprétations possibles, suivant que les élèves en question apprennent tous la même langue ou que la langue apprise dépend de chacun d'eux, ce qui se traduit respectivement par les formules logiques suivantes :
$\exists x[langue'(x) \wedge \forall y[eleve'(y) \longrightarrow apprend'(y,x)]]$
$\forall y[eleve'(y) \longrightarrow [\exists x[langue'(x) \wedge apprend'(y,x)]]$
La solution de Montague était de faire en sorte, encore une fois, que les phrases de ce genre aient *deux analyses syntaxiques différentes*, chacune donnant lieu à l'une des deux traductions sémantique possibles. Cette solution est difficilement applicable aux grammaires catégorielles de type AB, à moins d'ajouter de nouvelles affectations de catégories ou des nouveaux schémas de règles comme dans les Combinatorial Categorial Grammars (Steedman, 1996). Mais dans les grammaires de Lambek ou la logique linéaire, les deux analyses et les deux traductions sont produites "naturellement" par le système de règles et l'isomorphisme de Curry-Howard. La correspondance entre les structures est ainsi préservée, mais c'est parfois au prix de phénomènes de d"'ambiguités inutiles" (*spurious ambiguities*), quand les formules produites par ces diverses analyses se révèlent en fait équivalentes sur le plan logique.

Certains travaux récents ont pris une route différente, qui évite ce piège. L'idée est initialement venue du versant sémantique. Plutôt que de chercher à expliciter toutes les traductions logiques possibles de phrases contenant plusieurs quantificateurs, on s'est ainsi avisé qu'il était préférable de s'en tenir à une représentation *sous-spécifiée* de leur sens, c'est-à-dire à une expression qui se contente de spécifier les sens possibles sans les énumérer, à l'aide d'un *système de contraintes* (Reyle, 1995a; Reyle, 1995b; Koller et al., 2000; Erk et al., 2002). Il a ensuite été montré qu'on pouvait engendrer une expression sous-spécifiée de manière compositionnelle à partir d'une *analyse syntaxique unique* de la phrase ambiguë initiale. Ce travail a été réalisé avec les grammaires de dépendance (Dikovsky, 2005), et est en cours d'élaboration en exploitant les structures de dérivation des LTAG (Joshi et al., ; Pogodalla, 2004). Il permet, dans le pire des cas, un gain exponentiel en complexité algorithmique.

Cette idée séduisante n'a pu être mise à profit dans nos travaux, qui se sont atta-

chés à des formalismes "classiques" pour lesquels l'équation "autant d'analyses syntaxiques différentes que de sens possibles" devait être satisfaite. On peut d'ailleurs remarquer que si les représentations sémantiques sous-spécifiées sont utiles d'un point de vue calculatoire, pour restreindre l'explosion combinatoire des interprétations possibles compatibles avec un énoncé donné, leur pertinence cognitive est plus douteuse. En effet, si les énoncés du langage peuvent être ambigus, les situations qu'ils décrivent le sont en général beaucoup moins. Il est donc difficile de supposer que l'esprit humain les produit spontanément pour représenter ce qu'il perçoit par ses sens, indépendamment de la médiation d'une langue naturelle. Elles ne constitueront donc pas une donnée d'entrée de nos algorithme d'apprentissage.

Nous n'avons pas non plus cherché à exploiter une compositionnalité qui opérerait au niveau du discours, comme cela se pratique en DRT.

En guise de conclusion à cette section, nous pouvons maintenant reprendre la "partie basse" du schéma de la Figure 2.1 (qui concerne les niveaux qui nous intéressent particulièrement), en transformant les éléments qui la constituaient en données formelles : en reprenant les exemples décrits dans Exemple 1 et 5.2.1 et en partie 3.3.1, cela donne la Figure 3.4, où les donnnées importantes sont en gras.

Au vu de tout ce qui a été évoqués au cours de cette section et de la précédente, on mesure mieux la part de simplification que représente le passage d'un schéma à l'autre. Les formalismes, tant syntaxiques que sémantiques, que nous employons dans nos algorithmes, ne sont sans doute que des approximations encore très grossières de celles manipulées par les homos sapiens.

Même d'un point de vue purement informatique, les choix effectués ne sont pas les plus "à jour" des recherches actuelles. Pour la modélisation syntaxique, on aurait par exemple pu choisir les grammaires LTAG (dont l'expressivité middly context-sensitive est plus grande que celle des grammaires catégorielles) ou encore le formalisme des Abstract Categorial Grammars, plus général et homogène que ceux que nous employons. Pour la sémantique aussi, il existe des langages de représentation des connaissances dont la puissance d'expression est meilleure que la logique des prédicat du premier ordre à laquelle nous nous sommes restreint. Mais nous étions tributaires des théories disponibles à l'époque où nos travaux ont commencé et surtout, les grammaires de types AB sont les premières pour lesquelles des résultats *d'apprenabilité* ont été publiés.

Il reste donc à expliquer ce que ces résultats signifient, en modélisant l'apprentissage lui-même. C'est l'objet de la section suivante.

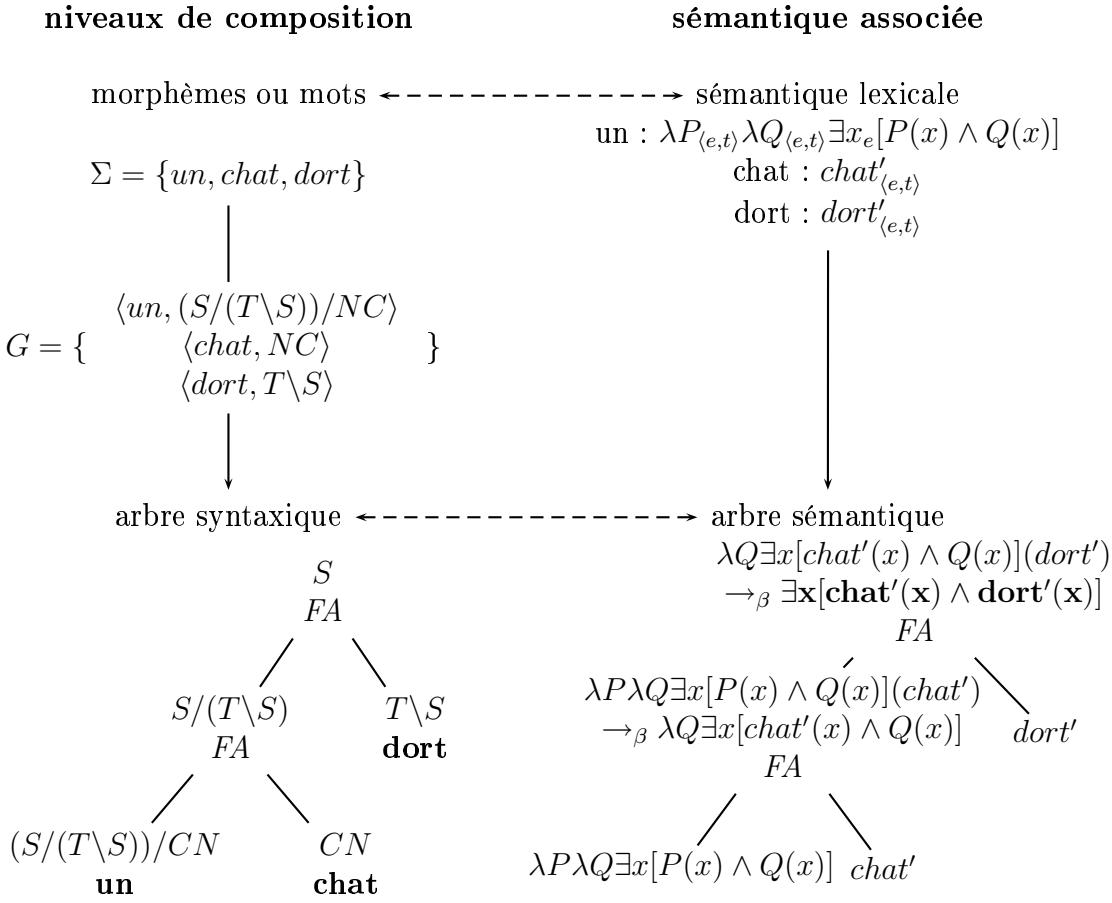

FIG. 3.4 – modélisation des niveaux d'analyse des langues naturelles

Chapitre 4

L'apprentissage (artificiel) et son langage

L'*apprentissage par induction*, c'est-à-dire la découverte de règles générales à partir de données qui les instancient, est un problème qui intéresse depuis longtemps les philosophes. Nombreux sont ceux parmi eux, depuis Hume jusqu'à Popper, qui ont douté de la possibilité même de telles opérations. Il y a souvent, en effet, une infinité d'explications différentes possibles compatibles avec un nombre fini d'observations. Popper, par exemple, préférait penser que les découvertes scientifiques sont le produit de raisonnements *déductifs* de type "essais/erreurs" (Popper, 1989).

En un sens, on peut dire que les recherches récentes en apprentissage artificiel[1] justifient l'intuition des philosophes. Elles montrent en particulier que l'induction est impossible à moins de disposer de connaissances *a priori* sur l'espace des hypothèses possibles, ce qu'on appelle un *biais*, (Mitchell, 1997; Cornuéjols & Miclet, 2002).

Dans cette partie, nous présentons d'abord les éléments qui constituent un problème d'apprentissage artificiel inductif en général, et le modèle dît "d'apprentissage à la limite" en particulier. Nous caractérisons le problème de l'acquisition de la syntaxe en ces termes. Puis, nous décrivons les résultats d'apprenabilité de classes de grammaires catégorielles qui ont été à l'origine de nos travaux. Ces classes, introduites dans la partie précédente, seront donc en quelque sorte réinterprétées ici comme un biais, au sens où elles caractérisent et contraignent l'espace de recherche des algorithmes d'apprentissage. Ce n'est que dans la partie suivante que les modèles sémantiques seront conjointement pris en compte dans les algorithmes d'apprentissage, mais nos choix de modélisation sont déjà évoqués ici.

[1] on traduit aussi souvent le terme anglais consacré de "machine learning" par "apprentissage automatique", mais nous préférons adopter ici la dénomination de (Cornuéjols & Miclet, 2002)

4.1 Critères et modèles d'apprentissage

L'apprentissage artificiel est, selon Mitchell, l'étude des programmes informatiques qui s'améliorent automatiquement avec l'expérience (Mitchell, 1997). Les enfants ne sont pas des programmes mais, à l'évidence, ils s'améliorent fortement avec l'expérience ! Nous passons en revue ici tout ce qu'on demande à un programme pour qu'il puisse être considéré comme un système d'apprentissage artificiel inductif.

4.1.1 Les composants de l'apprentissage artificiel inductif

Le domaine de l'apprentissage artificiel a une longue histoire, et couvre un champ de plus en plus vaste au fur et à mesure de l'amélioration de ses concepts et de ses techniques (Mitchell, 1997; Cornuéjols & Miclet, 2002). Nous nous restreignons ici à l'*apprentissage symbolique supervisé*, plus précisément à l'*induction à partir d'exemples*, dans le cas où l'on dispose d'une *connaissance explicite du domaine*, qui est un *espace structuré*. Nous détaillons dans ce qui suit ce que cela signifie.

Cette manière de considérer l'apprentissage artificiel remonte au moins à la fin des années 60, au commencement de l'âge d'or de l'intelligence artificielle "symbolique", celle qui cherchait à modéliser les processus psychologiques humains par le biais de techniques de *représentation des connaissances*. On cite souvent comme pionnier en la matière le programme "Arch" de Winston (Winston, 1995), qui date de cette époque : le système cherchait à apprendre la définition du concept d'"arches" construites dans un monde de blocs, à partir d'exemples et de contre-exemples présentés sous la forme de *réseaux sémantiques*.

Pour représenter le scénario classique de l'apprentissage par induction, Cornuejols et Miclet (Cornuéjols & Miclet, 2002) proposent un schéma que nous reproduisons (légèrement modifié) en Figure 4.1.

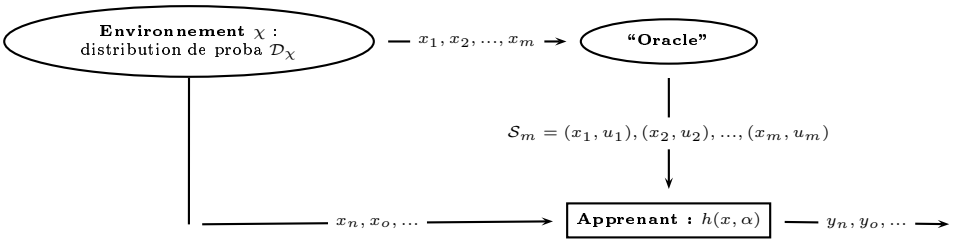

FIG. 4.1 – scénario fondamental de l'induction (d'après (Cornuéjols & Miclet, 2002))

Reprenons en détail les composants de ce schéma, pour énumérer ce qui spécifie un problème d'apprentissage artificiel inductif :

- Les *données d'apprentissage* $x_1,\ x_2,\ ..., x_n \in \chi$ sont tirées aléatoirement de l'environnement χ selon une distribution $\mathcal{D}_\chi$, fixe au cours du temps, sur l'espace d'entrée χ. Un "oracle" étiquette ces données en utilisant une fonction $f \in \mathcal{F}$: pour tout $i = 1, ..., n,\ u_i = f(x_i)$.
- La *cible* de l'apprentissage est la fonction f. A défaut de trouver exactement f, l'apprenant cherche à en obtenir une approximation h. Pour pouvoir chercher cette cible, il faut en fixer une *représentation* dans un *espace de recherche.*
- L'apprenant dispose de *connaissances initiales*, innées ou provenant d'acquisitions préalables, notées α sur le schéma : ces connaissances peuvent être de diverses natures, mais on peut y faire figurer l'espace de fonctions $\mathcal{H}$ à l'intérieur duquel h est recherché : c'est le fameux *biais* sans lequel aucun apprentissage n'est possible. On n'a pas nécessairement $f \in \mathcal{H}$.

Sur le schéma, on voit aussi les *données de test* $x_n, x_o, ...$, en principe différentes des données d'apprentissage, mais provenant du même environnement χ, auxquelles l'apprenant doit attribuer lui-même une étiquette $y_i = h(x_i)$. Mais pour définir rigoureusement un *modèle d'apprentissage*, il manque encore les composants suivants (qui ne sont pas visibles sur le schéma) :

- Le *protocole* suivi lors de l'apprentissage régit en particulier l'*évolution temporelle* du système : comment sont fournies les données d'apprentissage et les données de test (toutes en même temps ou l'une après l'autre...), à quel moment teste-t-on les réponses de l'apprenant ?
- Le dispositif doit aussi être muni d'un *critère d'arrêt* et/ou d'une *procédure d'évaluation* permettant d'établir le diagnostic de succès ou d'échec de l'apprentissage.

Une des premières idées que l'on peut avoir pour définir un apprenant est de mimer l'apprentissage "par coeur" en mémorisant simplement les données d'apprentissage au fur et à mesure qu'elles sont fournies. Cette technique, connue sous le nom de "memory-based learning" est mise en oeuvre dans certains systèmes opérationnels, y compris dans le domaine linguistique (Daelemans, 2005). Les données de test sont alors en général étiquetées par l'apprenant en recherchant parmi les données d'apprentissage celles qui s'en "rapprochent le plus", selon une notion de distance à préciser.

L'apprentissage statistique s'est aussi beaucoup développé ces dernières années dans le domaine de l'ingénierie linguistique. Dans ce cas, ce sont des propriétés de *fréquence* extraites des données d'apprentissage, qui servent à étiqueter les données de test. Les données d'apprentissage utilisées dans ce cas sont des *corpus de textes* plus ou moins enrichis (avec des étiquettes lexicales, voire même des indications

de structures comme dans les corpus arborescents), qui se veulent significatifs des phénomènes linguistiques étudiés. Peu de travaux encore, néanmoins, se sont consacrés à l'apprentissage sur corpus de grammaires catégorielles. Citons pourtant (Osborne & Briscoe, 1998) et (Steedman & alii, 2002) parmi les pistes les plus prometteuses, et (Moot, 2003) pour l'apprentissage par corpus d'une grammaire de type logique.

Mais l'approche que nous avons privilégiée est *symbolique* au sens où la cible de l'apprentissage sera élément d'un espace de recherche $\mathcal{H}$ structuré et où le critère de réussite sera une identification exacte de cette cible. L'apprentissage ainsi conçu a pour objectif de passer d'une *description extensionnelle* (un ensemble de données) à une *description intensionnelle* (un concept).

Comment cette grille d'analyse générale va-t-elle s'instancier dans notre problème d'apprentissage particulier ? C'est le moment de trouver des équivalents formels aux *conditions de l'apprentissage d'une langue chez les enfants* identifiées en partie 2.2.2.

4.1.2 Premiers choix de modélisation

Quelles sont les données dont dispose l'enfant pour son apprentissage ? C'est sur ce point que réside la principale originalité de nos recherches. Puisque, comme cela a été dit en partie 2.2.2, *l'accès au sens* est une condition nécessaire de l'acquisition d'une langue, nous faisons à l'origine de nos travaux l'hypothèse d'une *double donnée d'apprentissage* constituée d'un *énoncé* et d'une *représentation de son sens*. L'énoncé sera fourni sous la forme d'une *succession de symboles (des mots)*, ce qui, bien sûr, constitue une idéalisation puisque que l'enfant, lui, n'a accès qu'à un *flux acoustique non segmenté*. Mais l'apprentissage de cette segmentation a lieu au cours de ses premiers mois de vie (de Boysson-Bardie, 1999), et elle semble pouvoir s'effectuer sur la base de simples perceptions de régularités, modélisables par des statistiques (Brent, 1996). Nous supposons donc cette première phase déjà dépassée quand commence l'acquisition de la syntaxe. Pour représenter le sens des énoncés, nous adopterons principalement deux techniques différentes : soit en fournissant une *formule logique propositionnelle*, soit en associant aux mots le *type de leur traduction sémantique*. Dans les deux cas, cela suppose aussi que l'apprenant dispose, avant de commencer son apprentissage, d'une *compétence sémantique* qui le rend capable de traduire les perceptions de son environnement en représentations formelles.

Rappelons aussi que les langues naturelles sont apprises sur la base d'*exemples positifs seuls*. L'apprentissage à partir d'exemples positifs seuls est notoirement plus difficile que l'apprentissage par exemples positifs et négatifs, parce qu'une *surgénéralisation* n'y sera jamais démentie par aucun exemple négatif (nous reviendrons sur ce point important plus tard). Notons aussi que le schéma général

de l'apprentissage inductif fait de l'apprenant un système *passif*, alors que *l'interaction dialoguée* figure dans les conditions nécessaires identifiées en partie 2.2.2. On pourrait arguer de cette différence pour justifier, malgré tout, la fourniture d'exemples négatifs, en estimant qu'ils simulent des productions incorrectes corrigées. Mais c'est une voie que nous n'avons pas empruntée, parce que l'enfant semble pouvoir se passer de ce type de feed-back. Il faudrait plutôt intégrer le schéma de l'apprentissage inductif dans un *modèle de dialogue* ou adapter le schéma de *l'apprentissage par renforcement* à l'acquisition de grammaires, ce qui, à notre connaissance, n'a jamais été proposé jusqu'à présent.

Il y a en fait deux façons différentes de faire entrer nos données dans le schéma de la Figure 4.1. Soit on considère que les x_i sont des *couples* "énoncé/représentation sémantique" que l'oracle ne fait que valider (puisqu'on n'admet que des exemples positifs), soit on considère que les x_i sont seulement des "énoncés" et que c'est l'oracle qui leur associe leur représentation sémantique $u_i = f(x_i)$. Formellement, cela revient au même pour les entrées de l'apprenant, mais *la nature de la cible change*, puisque les données d'entrée et de sortie de h sont différentes dans les deux cas. C'est en fait cette deuxième solution que nous avons adoptée. Ainsi, mis en présence de nouveaux énoncés (les données de test $x_n, x_o, ...$), l'apprenant devra non seulement être capable de juger de leur grammaticalité, mais aussi leur associer une représentation sémantique $h(x_n), h(x_o), ...$ Il *apprend à comprendre.*

La cible de l'apprentissage est ainsi double : elle se compose d'une syntaxe et d'une fonction de traduction syntaxico-sémantique. Pour cette dernière, nous étudierons plusieurs variantes. La partie syntaxique, elle, sera toujours *représentée* par une grammaire formelle appartenant à une sous-famille spécifiée de grammaires logiques lexicalisées telles que celles définies en partie 3.1.3. Remarquons que, même si on admet que la compétence syntaxique humaine est représentée par une telle grammaire, il n'est pas certain qu'elle soit d'emblée la cible de l'apprentissage des enfants. On pourrait par exemple considérer que les différentes étapes d'acquisition de la langue définissent autant de "grammaires" différentes, servant de cibles successives (cf. la "grammaire archétypale" de Dikovsky (Dikovsky, 2004)).

Notons aussi que ce n'est pas parce que la grammaire cible est unique qu'elle correspond forcément à *la syntaxe correcte d'une langue officielle.* Notre approche n'est pas prescriptive : la grammaire à apprendre est celle pratiquée par l'environnement familial ou éducatif (s'il est constitué de plusieurs personnes, on suppose donc qu'elles emploient toutes la même), et non une norme fixée par on ne sait quelle instance légitime extérieure.

Au chapitre des connaissances initiales figurera, bien sûr, celle de la classe de grammaires à l'intérieur de laquelle doit s'effectuer la recherche de la grammaire cible (on la notera pour cela plutôt $\mathcal{G}$ par la suite), ainsi qu'une classe à laquelle appartient la fonction de traduction. Cette dernière devra en outre vérifier un

certain nombre de *conditions*, qui instancieront une version plus ou moins adaptée du *Principe de compositionnalité*, modélisé tel qu'en partie 3.3. Est-ce à dire que ce Principe est inné ou qu'il est, lui aussi, acquis très tôt ? Dans l'hypothèse innéiste, ce serait notre seule concession à la théorie chomskienne. Mais l'autre solution n'est pas forcément absurde non plus. Les expériences de (Kirby, 2002), qui montrent comment la compositionnalité émerge d'intéractions entre agents, peuvent être citées à l'appui des deux hypothèses, même si elles se situent plutôt dans la temporalité de la phylogénèse que dans celle de l'ontogénèse. Enfin, rappelons que le processus que nous souhaitons modéliser prend effet vers 2 ans, *après la mise en place de la sémantique lexicale*. Pour cette phase aussi, des simulations ont d'ores et déjà été proposées (Siskind, 1996; Kanazawa, 2001). Le stage de DEA en "sciences cognitives" de Cédric Messaint à Lille3, co-encadré en 2005 avec la linguiste Georgette Dal, a porté sur une meilleure compréhension des procédures employées par ces auteurs (Messiant, 2005). L'utilisation d'associations mot/représentation sémantique sera donc aussi permise dans nos systèmes.

Dans les 70 et 80, un certain nombre de travaux pionniers ont adopté des modélisations très similaires à celle-ci pour simuler l'apprentissage de la compétence syntaxico-sémantique, bien que la syntaxe y soit représentée par des grammaires formelles légèrement différentes (réseaux de transition récursif dans (Anderson, 1977), règles logiques dans (Langlet, 1982), grammaires transformationnelles dans (Hamburger & Wexler, 1975). Plus récemment (Feldman, 1998; Thompson et al., 1997) présentent également des systèmes capables d'acquérir des règles à partir de données syntaxiques et sémantiques. Mais, à nos yeux, il manque à ces recherches plus empiriques que théoriques une formulation explicite du Principe de compositionnalité. Seul Anderson doit faire appel à un principe de "non croisement de branches" entre les arbres syntaxique et sémantique, qui peut être vu comme une "version light" de la compositionnalité. De plus, ces travaux ne rentrent pas dans le cadre du schéma inductif détaillé dans la partie précédente, il faudrait leur associer un protocole et un critère de succès qui ne soit pas *ad hoc*. C'est sur ce point que nous nous penchons dans la section suivante.

4.1.3 Modèle d'apprentissage "à la limite"

Evaluer précisément la qualité d'un système d'apprentissage opérant à partir de phrases (d'énoncés), et dont la cible est une grammaire formelle n'est pas si simple qu'il y paraît. Ainsi, on comprend aisément que si la grammaire inférée par l'apprenant est formellement différente de la grammaire cible mais reconnaît le même langage, l'apprentissage doit être considéré comme réussi. Mais comparer les langages reconnus par deux grammaires différentes donne lieu à beaucoup de problèmes indécidables.

Un des premiers "modèle d'apprentissage" proposé adapté à cette situation

est celui d"'identification à la limite par exemples positifs seuls" de Gold (Gold, 1967). Même s'il est loin d'être totalement satisfaisant (cf. critiques dans la section suivante), c'est sur lui que nous nous sommes fondé. Nous le présentons ici dans sa version de base, où les données sont des phrases (nous verrons plus tard comment l'adapter à nos propres types de données). Soit donc $\mathcal{G}$ un ensemble de grammaires sur un alphabet Σ et soit la fonction $L : \mathcal{G} \longrightarrow pow(\Sigma^*)$ qui associe le langage $L(G)$ à chaque grammaire $G \in \mathcal{G}$ (cf. partie 3.1.3).

Definition 1 (Modèle de Gold par exemples positifs (Gold, 1967)).

- *Soit ϕ une fonction qui à tout échantillon fini de phrases de Σ^* associe une grammaire de $\mathcal{G}$ ou s'abstient. On dit que cette fonction* ***converge*** *vers $G \in \mathcal{G}$ sur un échantillon $\langle s_i \rangle_{i \in \mathbb{N}}$ d'éléments de Σ^* si $G_i = \phi(\langle s_0, ..., s_i \rangle)$ est défini et égal à G partout sauf pour un nombre fini de valeurs de $i \in \mathbb{N}$ ou, ce qui revient au même, si $\exists n_0 \in \mathbb{N}$ tel que pour tout $i \geq n_0$, G_i et défini et égal à G.*
- *On dit que ϕ* ***apprend*** *la classe $\mathcal{G}$ par exemples positifs si pour tout langage L de $L(\mathcal{G}) = \{L(G) | G \in \mathcal{G}\}$ et pour toute séquence $\langle s_i \rangle_{i \in \mathbb{N}}$ qui énumère L, c'est-à-dire telle que $\{s_i | i \in \mathbb{N}\} = L$, il existe $G \in \mathcal{G}$ telle que $L = L(G)$ et ϕ converge vers G sur $\langle s_i \rangle_{i \in \mathbb{N}}$.*
- *$\mathcal{G}$ est* ***apprenable*** *s'il existe une fonction ϕ calculable qui apprend $\mathcal{G}$.*

Le premier point de cette définition fixe le protocole et le critère d'arrêt. Le temps est implicitement discrétisé par $\mathbb{N}$ et l'apprenant, représenté par la fonction ϕ, reçoit une nouvelle donnée s_i (c'est-à-dire une phrase, un énoncé) à chaque unité de temps i. A chaque réception, il s'abstient ou formule une hypothèse G_i (qui est une grammaire), en tenant compte de toutes les données qu'il a reçues jusqu'à présent : $G_i = \phi(\langle s_0, ..., s_i \rangle)$. Le "critère d'arrêt" est, en l'occurrence, un peu particulier puisque, justement, il n'impose pas d'arrêt définitif et s'inspire plutôt de la notion de "limite" en mathématiques. En observant les sorties successives G_0, G_1, G_2, ... de ϕ, on ne sait jamais si une "grammaire limite" est atteinte ou non. Gold mettait en avant l'apprentissage "tout au long de la vie" pour justifier cette définition. Personne, en effet, ne peut prétendre que sa propre compétence syntaxique est fixe et définitive, même si l'essentiel est acquis dans les premières années de vie.

Le deuxième point est le *critère de succès*. Il est important de remarquer qu'il porte non sur une grammaire particulière mais sur une *classe de grammaires*. L'apprenant n'est pas spécialiste dans l'apprentissage d'une grammaire particulière, mais il peut apprendre toutes celles appartenant à une certaine famille. C'est évidemment un point crucial pour la crédibilité psychologique du modèle. Un bébé humain est capable d'apprendre n'importe laquelle des quelque 5 000 langues naturelles identifiées dans le monde, pourvu qu'il soit mis dans un environnement adapté.

Une grammaire étant fixée, le modèle de Gold ne requiert pas la définition d'une distribution de probabilité sur les données fournies à l'apprenant ; il demande juste que l'algorithme se comporte correctement sur tout échantillon de données qui, à la limite, énumère le langage de cette grammaire. Notons aussi que, comme attendu, si ϕ converge vers une grammaire qui n'est pas la même que la grammaire cible, mais génère exactement le même langage, l'apprentissage est un succès. Nous verrons plus loin comment éviter d'avoir à réaliser des opérations indécidables, comme de tester l'inclusion entre langages algébriques. En effet, le dernier point de la définition stipule que l'apprenant doit être un programme.

4.1.4 Résultats classiques, intérêts et limites

Dans son article fondateur de 1967, Gold ne se contentait pas de proposer son modèle d'apprentissage[2] ; il démontrait aussi les premiers théorèmes auquel il donnait lieu. Le plus important pour nous est malheureusement un résultat négatif. Il stipule que tout ensemble de grammaires capable de générer *tous les langages finis et au moins un langage infini* est non apprenable à la limite par exemples positifs. Essayons de donner l'intuition sous-jacente à ce résultat, en supposant que la classe de grammaires à apprendre génère tous les langages finis plus le langage Σ^*. Quelle que soit la stratégie adoptée par l'apprenant pour identifier une grammaire de cette famille, elle peut être mise en échec. En effet[3] :

- soit la stratégie consiste à proposer systématiquement comme hypothèse une grammaire générant un langage fini qui contient tous les exemples proposés jusqu'à présent : cette stratégie échouera toujours à identifier la grammaire générant Σ^* et à converger vers elle ;
- soit la stratégie aboutit parfois à converger vers une grammaire générant Σ^* (par définition, toujours compatible avec les données), auquel cas elle prend le risque de sur-généraliser et donc d'échouer si la cible ne générait qu'un langage fini.

Corollairement, ni l'ensemble des grammaires algébriques, ni non plus celui des grammaires catégorielles de type AB (cf. partie 3.1.3) qui génère la même classe de langages, ni même celui des grammaires régulières ne sont apprenables. Même si les arguments avancés dans cette preuve sont difficiles à interpréter en termes psycholinguistiques, ils sont rédhibitoires pour l'apprenabilité "à la limite" des classes les plus simples de la théorie des langages.

[2] il en proposait en fait plusieurs, nous n'évoquons ici que celui qui est adapté à notre modélisation

[3] pour simplifier, nous n'envisageons ici que des stratégies *consistantes*, c'est-à-dire ne produisant que des hypothèses compatibles avec l'ensemble des données reçues ; cette condition n'est pas requise par le modèle de Gold

Ce résultat a longtemps semblé sonner le glas des recherches sur le sujet. Dans son article fondateur pourtant, Gold évoquait déjà des pistes pour continuer malgré tout ses travaux. Il évoquait notamment la possibilité de considérer des classes de grammaires différentes de celles figurant dans la hiérarchie de Chomsky. Rien ne s'oppose en effet à ce qu'un ensemble de grammaires, générant un nombre infini de langages distincts mais pas *tous* les langages finis, puisse être apprenable. Peut-être que les langues naturelles constituent une famille de ce genre.

Ce n'est que dans les années 80, avec notamment les travaux d'Angluin (Angluin, 1992), que les contours de telles classes ont commencé à apparaître (nous en évoquons certaines en partie 4.2). Un des principaux apports d'Angluin est la caractérisation des familles (énumérables) de langages récursifs telles que les classes de grammaires qui les engendrent[4] sont apprenables à la limite (Angluin, 1980b; Angluin, 1980a). Cette caractérisation fait appel à la notion d'ensembles caractéristiques, et s'accompagne d'un schéma d'algorithme d'apprentissage qui satisfait les conditions de Gold. Elle a permis la démonstration de preuves d'apprenabilité (ou de non apprenabilité) plus simples.

L'autre piste plus récemment explorée pour passer outre le résultat de non-apprenabilité de Gold est celui d'une *redéfinition de la notion de langage*, qui va avec un *enrichissement des données d'entrée*. Ainsi, pour apprendre des grammaires algébriques (Sakakibara, 1990; Sakakibara, 1992) ou des grammaires catégorielles (Kanazawa, 1996; Kanazawa, 1998), il est plus facile de disposer de *données structurées*, qui rendent compte de la nature arborescente des analyses syntaxiques qu'elles produisent. Nous revenons dans la partie suivante (cf partie 4.2) sur cette approche. Le modèle de Gold donné en Définition 1 est alors adapté en remplaçant Σ^* par un nouvel espace plus riche, et en redéfinissant la fonction L en conséquence, pour qu'elle prenne ses valeurs dans ce nouvel espace. Par exemple, comme nous le verrons, L peut désigner une notion de "langage de structures". Bien sûr, même dans ce cas, la preuve de non apprenabilité esquissée précédemment peut être adaptée : toute famille de grammaires engendrant *tous les langages de structures finis et au moins un lagage de struture infini* n'est pas apprenable à la limite. Cette approche s'emploie donc *conjointement* à une restriction de la classe des grammaires considérées.

Au fil des années, plusieurs conditions suffisantes d'apprenabilité ou de non apprenabilité concernant des classes de grammaires sont venues affiner la caractérisation d'Angluin. Ces conditions portent encore sur les langages engendrés par ces classes. Le schéma de la figure 4.2, inspiré de (Florêncio, 2003) (mais simplifié), explicite les relations entre certaines de ces conditions. Nous ne faisons figurer sur

[4]pour toutes les conditions portant sur des classes de langages, on fait l'hypothèse que le "membership problem", c'est-à-dire l'appartenance d'une phrase au langage d'une grammaire donnée, est décidable

ce schéma que les critères les plus utiles, qui seront repris et explicités plus loin, en partie 4.2.3. Cette *hiérarchie d'apprenabilité* n'entretient pas de liens simples avec la hiérarchie de Chomsky : l'expressivité et l'apprenabilité sont des propriétés bien distinctes.

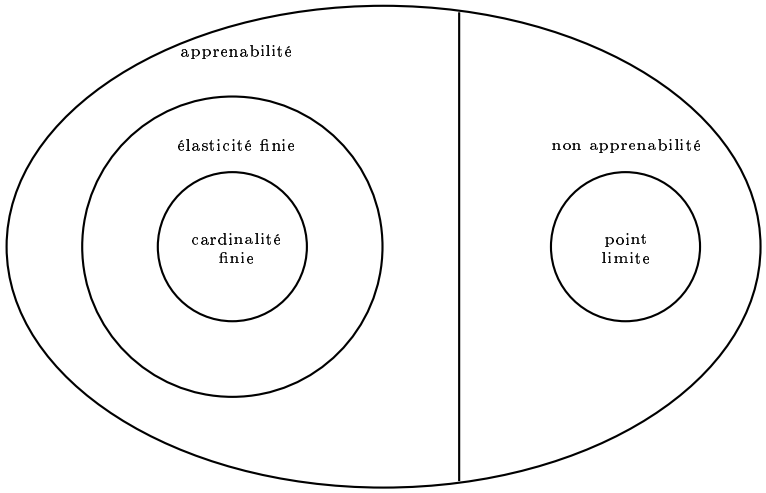

FIG. 4.2 – hiérarchie de classes de langages suivant le critère d'apprenabilité

Même si ces résultats laissent entendre que la notion d'apprenabilité de Gold n'est pas aussi grossière qu'on aurait pu le craindre au premier abord, de nombreuses critiques ont été émises à son encontre. On peut ainsi regretter que le critère de succès n'impose aucune condition de complexité sur l'algorithme d'apprentissage. Pitt (Pitt, 1989) a montré qu'il est difficile de contraindre la complexité de l'apprenant, dans la mesure où il dispose d'un temps potentiellement infini pour converger. Le fait de l'obliger à identifier une grammaire qui reconnaît *exactement* le même langage que celui dont sont issues les données est également contestable. Un tel dispositif est évidemment complètement hermétique à toutes *données bruitées*, ce qui est peu réaliste.

De nombreuses variantes au modèle de Gold ont été proposées, en général en posant des contraintes sur les propriétés de son apprenant. Des alternatives plus radicales ont aussi été avancées. La plus connue est sans doute l'apprentissage PAC ("Probablement Approximativement Correct") de Valiant (Valiant, 1984). Dans ce modèle, la classe des grammaires régulières n'est toujours pas apprenable par exemples positifs. Toutefois, des variantes plus restrictives de PAC rendent cet apprentissage possible. (Denis et al., 1996; Denis & Gilleron, 2001) ont ainsi montré qu'en posant des contraintes simples sur les distributions de probabilité

acceptables dans le modèle, l'apprentissage des grammaires régulières devenait possible.

D'autres approches encore proposent de modéliser le rapport enseignant/enseigné, en autorisant des intéractions entre eux. Le modèle par *requêtes* d'Angluin autorise l'apprenant à poser certaines questions explicites au "professeur" qui lui fournit des exemples (Angluin, 1987; Angluin, 1988). Le modèle de Goldman et Mathias (Goldman & Mathias, 1996), adapté aux langages par La Higuera (de la Higuera, 1997), met en scène un enseignant et un "adversaire" sensé éviter la collusion entre lui et l'apprenant. Mais les langues naturelles, comme nous l'avons vue, sont acquises sans "leçons" explicites.

Au moment où nous avons commencé nos travaux, le modèle de Gold connaissait un certain regain d'intérêt, dû principalement aux résultats d'apprenabilité de Kanazawa (Kanazawa, 1996), que nous présentons en détail dans la section suivante. Comme ces résultats concernaient des classes de grammaires catégorielles, ils ont naturellement constitué une des bases de nos propres recherches. Nous nous en sommes donc tenu à ce modèle, qui s'est avéré suffisamment souple pour être facilement adaptable au type de données syntaxico-sémantiques dont nous avions besoin.

4.2 L'apprentissage automatique de grammaires catégorielles

Le critère d'apprenabilité "à la limite" de Gold impose l'existence d'un algorithme d'apprentissage mais dit peu de choses sur son fonctionnement. Avec les conditions nécessaires et/ou suffisantes de la Figure 4.2, il est même possible de démontrer l'apprenabilité d'une classe de grammaires sans exhiber pour autant un algorithme d'apprentissage. Ceux décrits par Gold dans son article fondateur procédaient à une énumération exhaustive de la classe à apprendre ; ce type de stratégie n'est évidemment ni efficace ni très pertinent d'un point de vue cognitif. L'algorithme "générique" d'Angluin pour l'apprentissage par exemples positifs est déjà nettement plus intéressant. Depuis, les travaux de Mitchell (Michalski et al., 1986) ont aussi contribué à reconsidérer l'apprentissage automatique symbolique comme un problème de recherche dans un espace structuré. Nous allons voir dans cette partie comment tirer parti de connaissances sur la structure de l'espace de recherche pour construire des algorithmes d'apprentissage aussi efficaces que possible. Nous l'illustrons d'abord en détail avec une sous-classe des grammaires catégorielles de type AB (notées GCs par la suite), dont les propriétés ont été étudiées par Buszkowki et Penn (Buszkowski & Penn, 1990) et dont l'apprenabilité a été analysée par Kanazawa (Kanazawa, 1998). Puis nous recensons les

principaux résultats connus depuis sur l'apprenabilité de classes de grammaires catégorielles.

4.2.1 Exemples structurés et treillis des GCs rigides

Nous avons évoqué dans la partie précédente qu'il était possible d'associer aux GCs un *langage* qui rende mieux compte de la nature arborescente de leurs analyses syntaxiques. C'est ce que nous faisons ici. Nous appelons tout d'abord *FA-structure* (FA pour "Foncteur-Argument") sur un vocabulaire Σ tout arbre binaire dont les feuilles sont étiquetées par des éléments de Σ et dont chaque noeud est étiqueté soit par BA soit par FA. L'ensemble des FA-structures sur Σ est noté Σ^F. Pour toute GC $G \subset \Sigma \times Cat(\mathcal{B})$, un *exemple structuré* pour G est un élément de Σ^F qui est obtenu à partir d'un arbre d'analyse syntaxique produit par G pour une phrase $w \in L(G)$, en effaçant dans cet arbre toutes les catégories de $Cat(\mathcal{B})$. Enfin, pour toute GC G, le *langage des structures* de G, noté $FL(G)$ est l'ensemble de ses exemples structurés. La Figure 4.3 présente deux FA-structures éléments du langage de structures de la grammaire définie dans l'Exemple 1, partie 3.1.3 (en choisissant pour "un" la catégorie $(S/(T\backslash S))/NC$, comme justifié par la compositionnalité en partie 3.3.1).

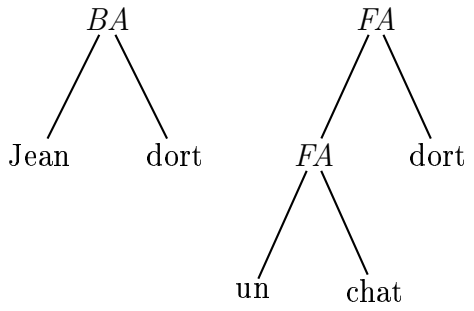

FIG. 4.3 – deux FA-structures éléments de $FL(G)$

Parallèlement au langage de structures, on définit une relation d'ordre partielle sur l'ensemble des grammaires catégorielles grâce à la notion de *substitution*. Soit χ un ensemble infini dénombrable de variables et soit $\mathcal{B} = \chi \cup \{S\}$. Une *substitution* est une fonction $\sigma : \chi \longrightarrow Cat(\mathcal{B})$ qui transforme une variable en une catégorie. Par convention, pour définir une substitution σ, on définit uniquement sa valeur sur les éléments de χ qu'elle modifie (par défaut, $\sigma(x) = x$). Une substitution est étendue par morphisme à une fonction de $Cat(\mathcal{B})$ dans $Cat(\mathcal{B})$ de la manière suivante : (i) $\sigma(S) = S$, (ii) $\sigma(A/B) = \sigma(A)/\sigma(B)$ et (iii) $\sigma(A\backslash B) = \sigma(A)\backslash\sigma(B)$ pour tout $A, B \in Cat(\mathcal{B})$. De même, une substitution peut être étendue à une grammaire catégorielle quelconque G : $\sigma(G) = \{\langle v, \sigma(A)\rangle | \langle v, A\rangle \in G\}$.

Une substitution σ est dite *fidèle* à une grammaire G si la condition suivante est vérifiée : pour tout $v \in \Sigma$, si $\langle v, A\rangle \in G$ et $\langle v, B\rangle \in G$ et $A \neq B$ alors $\sigma(A) \neq \sigma(B)$. Soit $\sqsubseteq$ la relation binaire sur les grammaires définie par : $G_1 \sqsubseteq G_2$ si et seulement si il existe une substitution σ fidèle à G_1 telle que $\sigma(G_1) \subseteq G_2$. Par une substitution fidèle, il est impossible *d'unifier* les catégories distinctes affectées à un même mot (cela aura peu d'incidences par la suite car cette notion sera surtout employée dans le domaine des GCs *rigides*, où chaque mot est associé à au plus une catégorie). En revanche, une telle substitution permet par exemple de transformer une catégorie de base en une catégorie qui, elle, n'est plus dans $\mathcal{B}$. $\sqsubseteq$ définit une relation d'ordre sur l'ensemble $\mathcal{G}$ des GCs.

Il nous manque quelques dernières définitions générales portant sur les GCs. Une GC est dite *sans catégorie inutile* si toutes ses affectations de catégories sont employées dans au moins une analyse syntaxique. Transformer une GC quelconque en une GC sans catégorie inutile et ayant le même *langage de structure* est un processus calculable. On peut le considérer comme une mise en forme normale de la grammaire initiale. Enfin, pour tout entier $k \geq 1$, l'ensemble des GCs qui affectent au plus k catégories distinctes à chacun des mots de leur vocabulaire est appelée la classe des *GCs k-valuées*, et est notée $\mathcal{G}_k$. Pour $k = 1$, les grammaires de $\mathcal{G}_1$ sont aussi dîtes *rigides*.

C'est sur cette classe que nous nous concentrons dans un premier temps. Buskowski et Penn ont démontré la propriété fondamentale suivante : soit G_1 et G_2 dans $\mathcal{G}_1$ avec G_1 sans catégorie inutile, alors $G_1 \sqsubseteq G_2$ si et seulement si $FL(G_1) \subseteq FL(G_2)$. Cette relation établit un lien entre l'espace des grammaires catégorielles de $\mathcal{G}_1$ et l'espace des *données structurées* qu'elles produisent. Ce lien est visualisé sur la Figure 4.4, où la relation d'ordre entre grammaires est représentée par une flèche. La relation $G_1 \sqsubseteq G_2$ signifie en quelque sorte que "la grammaire G_2 est plus générale que la grammaire G_1" au sens où elle produit un langage de structures plus grand. L'application d'une substitution sur une grammaire est une opération de *généralisation*.

Kanazawa (Kanazawa, 1996) a montré que l'ensemble $\mathcal{G}_1$ auquel on ajoute un élément "top" noté $\top$, plus grand que tous les autres au sens de la relation $\sqsubseteq$, est un *treillis complet* pour cette relation d'ordre, c'est-à-dire que tout sous-ensemble de $\mathcal{G}_1 \cup \{\top\}$ a une borne supérieure et une borne inférieure dans $\mathcal{G}_1 \cup \{\top\}$[5]. De plus, on peut facilement calculer explicitement la borne supérieure d'un ensemble de grammaires $G_1, G_2, ..., G_n$ dans $\mathcal{G}_1 \cup \{\top\}$. Cette grammaire, notée $G_1 \sqcup G_2 ... \sqcup G_n$ s'obtient en cherchant l'*unificateur le plus général* (ou mgu) de $\{G_1, G_2, ..., G_n\}$, c'est-à-dire la plus petite substitution qui, appliquée à chacune des grammaires de cet ensemble, donne la même grammaire de $\mathcal{G}_1$. Si elle existe, cette grammaire est exactement (à un renommage près des catégories de base) $G_1 \sqcup G_2 ... \sqcup G_n$, sinon

[5]la grammaire de $\mathcal{G}_1$ qui joue le rôle de $\bot$ est $\emptyset$

$G_1 \sqcup G_2 ... \sqcup G_n = \top$.

Nous allons maintenant voir comment exploiter les propriétés de cette structure pour obtenir un algorithme d'apprentissage à partir d'exemples structurés positifs qui soit valable sur la classe $\mathcal{G}_1$. C'est la démarche qu'a réalisée Kanazawa, dont nous exposons dans la partie suivante les grandes lignes, sans donner les démonstrations complètes mais en en présentant les points clés.

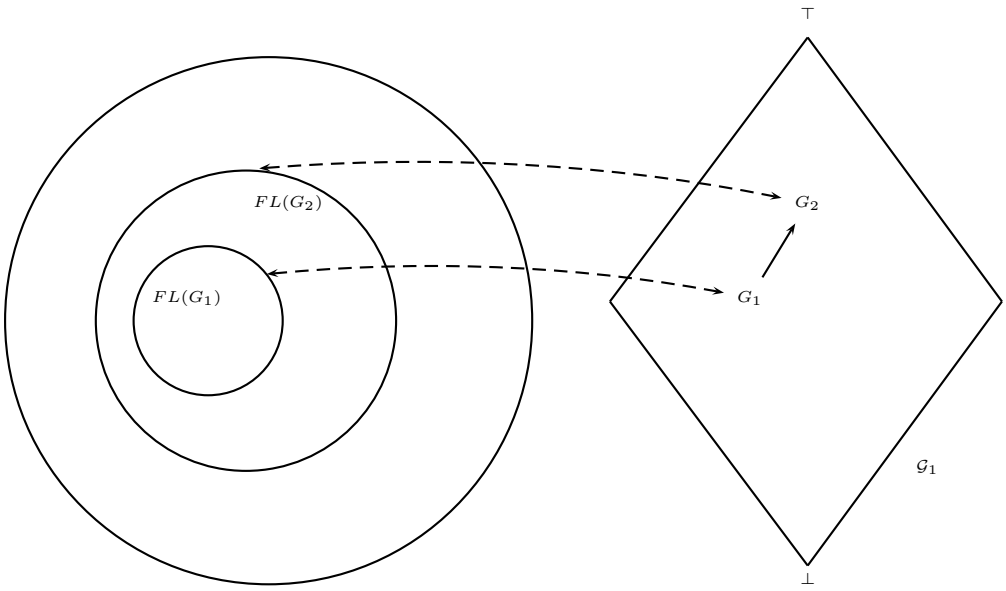

FIG. 4.4 – lien entre langages de structures et ordre sur les grammaires

4.2.2 Apprentissage de $\mathcal{G}_1$ par généralisation

L'algorithme RG (pour "Rigid Grammars") dû à Buszkowki et Penn, prend en entrée un ensemble D de FA-structures et donne comme résultat, si elle existe, une grammaire de $\mathcal{G}_1$ compatible avec cet ensemble D. L'algorithme est le suivant :

1. A partir de D, on construit tout d'abord une GC particulière appelée "forme générale de D" et notée $FG(D)$ dont le langage de structure $FL(FG(D))$ est *exactement* D. Pour cela, il faut :
   - introduire l'étiquette S à la racine de chaque exemple structuré ;
   - introduire une variable distincte x_i à chaque noeud argument de D (c'est-à-dire, au fils gauche de chaque noeud étiqueté BA, et au fils droit de chaque noeud FA) ;
   - introduire à chaque noeud restant la catégorie qui rend possible l'application des schémas FA et BA qui étiquettent ces noeuds.

- récolter les catégories affectées aux feuilles des arbres à l'issue des étapes précédentes, et les associer aux mots figurant à ces mêmes feuilles.

2. si $FG(D) \in \mathcal{G}_1$ alors $RG(D) = FG(D)$. Sinon, pour chaque mot du vocabulaire auquel $FG(D)$ affecte plus d'une catégorie, chercher l'unificateur le plus général de cet ensemble de catégories. S'il existe, il définit une substitution σ. Le résultat de l'algorithme est alors $RG(D) = \sigma(FG(D))$.

Exemple 3 (Calcul de $RG(D)$). *La Figure 4.5 montre le résultat de la première étape de RG appliqué aux exemples structurés de la Figure 4.3. $FG(D)$ est donc définie par : $FG(D) = \{\langle Jean, x_1\rangle, \langle dort, x_1\backslash S\rangle, \langle dort, x_2\rangle, \langle chat, x_3\rangle, \langle un, (S/x_2)/x_3\rangle\}$. Dans notre cas, $FG(D) \notin \mathcal{G}_1$ à cause des deux catégories distinctes affectées au mot "dort". L'unificateur le plus général de ces deux catégories est obtenue par la substitution : $\sigma(x_2) = x_1\backslash S$ (et σ est l'identité partout ailleurs). On a donc finalement : $RG(D) = \{\langle Jean, x_1\rangle, \langle dort, x_1\backslash S\rangle, \langle chat, x_3\rangle, \langle un, (S/(x_1\backslash S))/x_3\rangle\}$.*

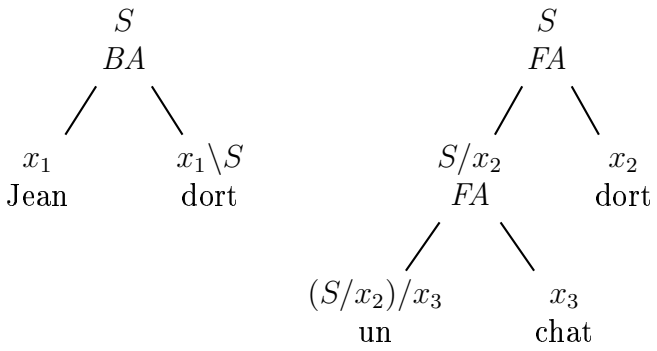

FIG. 4.5 – construction de $FG(D)$

$RG(D) = \sigma(FG(D))$ est sans catégorie inutile, et on a $D \subseteq FL(RG(D))$. Mais la propriété fondamentale de la grammaire $RG(D)$ ainsi construite, si elle existe, est qu'elle coïncide exactement (au renommage des catégories de base près) avec la borne inférieure de l'ensemble $\{G \in \mathcal{G}_1 | D \subseteq FL(G)\}$, c'est-à-dire de l'ensemble des GCs rigides produisant D, quand celle-ci est différente de $\perp = \emptyset$. $RG(D)$ est en quelque sorte le "moindre généralisé" dans $\mathcal{G}_1$ de l'ensemble des GCs quelconques compatibles avec l'ensemble initial D. On visualise cette situation sur la Figure 4.6. Notons aussi que l'algorithme RG est incrémental, au sens où $RG(D_1 \cup D_2) = RG(D_1) \sqcup RG(D_2)$ pour tous ensembles de FA-structures D_1 et D_2.

Il reste à s'assurer que l'algorithme RG satisfait le critère d"'apprentissage à la limite" de Gold sur l'espace $\mathcal{G}_1$ pour la notion de langage de structures. Cela revient à s'assurer que pour tout $G \in \mathcal{G}_1$ et toute suite $\langle s_i\rangle_{i\in\mathbb{N}}$ de FA-structures qui

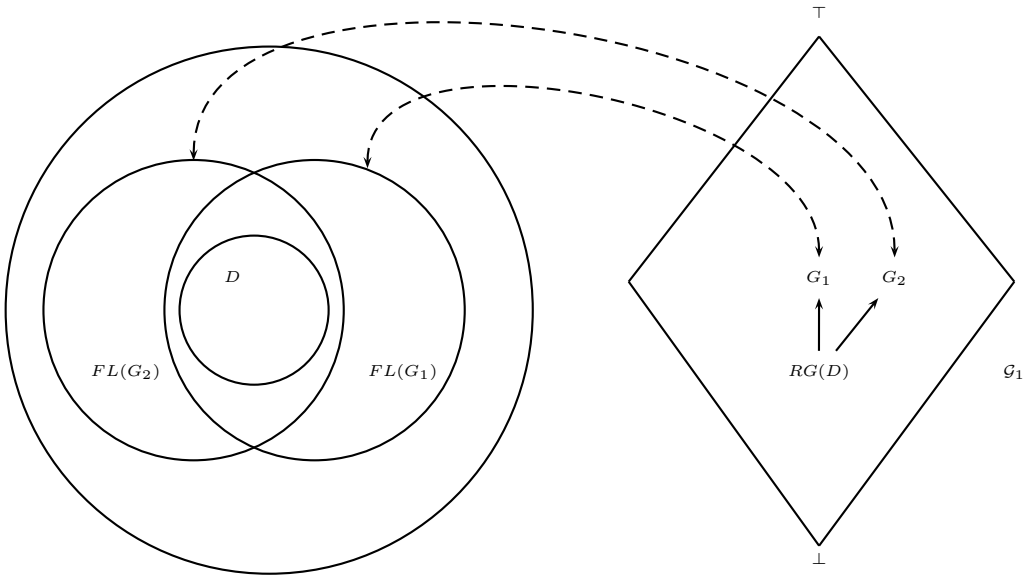

FIG. 4.6 – construction de la grammaire $RG(D)$

énumère $FL(G)$, $RG(\{s_0, s_1, ..., s_n\})$ *converge* vers une grammaire G' qui produit le même langage de structures que G : $FL(G') = FL(G)$.

C'est là que l'étude de l'ensemble $FL(\mathcal{G}_1) = \{FL(G)|G \in \mathcal{G}_1)\}$ va jouer un rôle déterminant. Kanazawa prouve en particulier que pour toute suite strictement ascendante $FL_1 \subset FL_2 \subset ... \subset FL_k$ de langages de structures sur un alphabet Σ dans $FL(\mathcal{G}_1)$, on a nécessairement $k \leq |\Sigma|$. Or, pour tout $n \in \mathbb{N}$, on a : $RG(\{s_0, s_1, ..., s_n\}) = RG(\{s_0\}) \sqcup RG(\{s_1\}) \sqcup ... \sqcup RG(\{s_n\})$ donc on a aussi : $RG(\{s_0\}) \sqsubseteq RG(\{s_0, s_1\}) \sqsubseteq ... \sqsubseteq RG(\{s_0, s_1, ..., s_n\})$. Ainsi, la suite des langages de structures générées par ces grammaires de plus en plus générales est ascendante dans $FL(\mathcal{G}_1)$: $FL(RG(\{s_0\})) \subseteq FL(RG(\{s_0, s_1\})) \subseteq ... \subseteq FL(RG(\{s_0, s_1, ..., s_n\}))$. D'après la propriété précédente, cette suite doit être stationnaire au bout d'un certain rang $N \in \mathbb{N}$, ce qui à son tour n'est possible que si $(RG(\{s_i\}_{i\leq n}))_{n\in\mathbb{N}}$ elle-même est stationnaire à partir de ce rang N. Il est facile, à partir de là, de se convaincre que la grammaire $RG(\{s_i\}_{i\leq N}) = G'$ génère le même langage de structures que la grammaire cible G. En effet :

- pour tout $n \in \mathbb{N}$, $RG(\{s_i\}_{i\leq n}) \sqsubseteq G$ par construction de $RG(\{s_i\}_{i\leq n})$ en tant que moindre généralisé donc pour $n = N$ on a $G' \sqsubseteq G$ et $FL(G') \subseteq FL(G)$;
- Pour tout $n \in \mathbb{N}$, $\{s_i\}_{i\leq n} \subseteq FL(RG(\{s_i\}_{i\leq n})) \subseteq FL(G')$ et $\{s_i\}_{i\in\mathbb{N}}$ énumère $FL(G)$ donc $FL(G) \subseteq FL(G')$.

Kanazawa a ainsi prouvé que l'ensemble $\mathcal{G}_1$ est apprenable par exemples structurés positifs par l'algorithme RG, qui en plus présente plusieurs "bonnes proprié-

tés" (il est efficace, consistant, incrémental, monotone...). Ce résultat fondamental est le point de départ de nombreux autres. Nous les évoquons maintenant, en rentrant un peu moins dans les détails.

4.2.3 Autres résultats d'apprentissage sur les grammaires catégorielles

La propriété sur les chaînes strictement ascendantes de langages de structures de $FL(\mathcal{G}_1)$ est en fait un cas particulier d'*élasticité finie*. Un ensemble $\mathcal{L}$ de langages possède la propriété d'*élasticité infinie* s'il existe une suite infinie $\langle s_n \rangle_{n\in\mathbb{N}}$ de phrases et une suite infinie $\langle L_n \rangle_{n\in\mathbb{N}}$ de langages de $\mathcal{L}$ telles que $\forall n \in \mathbb{N} : s_n \notin L_n$ et $\{s_0, ..., s_n\} \subseteq L_{n+1}$. L'élasticité finie est la négation de l'élasticité infinie, elle constitue une condition suffisante d'apprenabilité pour les familles de grammaires qui engendrent de telles classes de langages (cf. Figure 4.2). Kanazawa (Kanazawa, 1996) a montré que si une classe $\mathcal{M}$ de langages sur un alphabet Γ est d'élasticité finie et si $R \subseteq \Sigma^* \times \Gamma^*$ est une relation finie, alors $\mathcal{L} = \{R^{-1}[M] | M \in \mathcal{M}\}$ a aussi une élasticité finie. Ce théorème permet d'étendre le résultat d'apprenabilité de la classe $\mathcal{G}_1$ par exemples structurés positifs a d'autres classes.

La première extension naturelle consiste à considérer les classes $\mathcal{G}_k$ de GCs k-valuées, c'est-à-dire associant au maximum k catégories différentes à chaque mot de leur vocabulaire ($k \geq 1$). Ces classes sont toutes également apprenables par exemples structurés positifs. Malheureusement, les ensembles $\mathcal{G}_k$ pour $k > 1$ ne forment pas une structure de treillis comme pour $\mathcal{G}_1$. L'algorithme d'apprentissage RG donné en partie 4.2.2 demande donc plusieurs adaptations : l'unification de catégories pour associer au maximum k catégories distinctes à chaque mot, au lieu de produire un résultat unique, construit un *ensemble de grammaires*. Parmi celles-ci, pour éviter les surgénéralisations, on ne garde que les grammaires qui génèrent le plus petit langage de structures (l'inclusion des langages de structures est décidable) et on en choisit une comme résultat. Cet algorithme converge à la limite, mais il n'est ni efficace ni incrémental. Il est douteux que d'autres approches puissent être plus performantes, puisque Costa-Florencio a montré que l'identification de l'ensemble des grammaires k-valuées consistantes avec un ensemble de FA-structures est un problème NP-complet dès que $k > 1$ (Florêncio, 2000; Florêncio, 2001).

Se ramener à des exemples non structurés est une autre façon d'étendre le résultat d'apprenabilité initial. Dans les grammaires de type AB, il existe en effet toujours un nombre fini d'exemples structurés possibles compatibles avec une phrase donnée. Dans toute classe $\mathcal{G}_k$, il est donc possible d'énumérer toutes ces structures possibles, et de leur appliquer ensuite l'algorithme d'apprentissage par exemples structurés évoqué précédemment. Un problème persiste, pour éviter la

surgénéralisation : l'inclusion des langages de chaînes $L(G_1) \subseteq L(G_2)$ entre deux GCs G_1 et G_2 est indécidable. Pour contourner le problème, Kanazawa propose de ne tester, à chaque unité de temps i auquel est appliqué l'algorithme, que l'inclusion entre phrases de longueurs au plus i (il y en a un nombre fini). Comme l'algorithme dispose d'un temps potentiellement infini dans le critère de Gold, ce test suffira à sélectionner "à la limite" les grammaires générant un langage minimal. Mais, là encore, l'efficacité n'est pas au rendez-vous. L'explosion combinatoire est sensible même pour les GCs rigides : Costa Florencio a montré que l'identification de l'ensemble des GCs de $\mathcal{G}_1$ compatibles avec un ensemble de phrases est NP-complet (Florêncio, 2002). L'apprentissage de $\mathcal{G}_k$, $k > 1$ à partir de chaînes est d'autant plus inaccessible.

Ces résultats ont été le point de départ de divers travaux fructueux. Les contributions que nous citons par la suite sont toutes issues de l'Action de Recherche Coopérative "Gracq" (pour "Grammar Acquisition") de l'Inria[6], qui a coordonné les efforts de plusieurs équipes françaises sur le sujet entre 2000 et 2002. Une bonne part de nos propres recherches, exposées dans la section suivante, s'y rattachent également.

Dans les travaux de Kanazawa, on l'a vu, seul l'apprentissage de la classe $\mathcal{G}_1$ donne lieu à un algorithme "praticable". Besombes et Marion (Besombes & Marion, 2004) ont défini une nouvelle classe, qui étend strictement $\mathcal{G}_1$ tout en ayant une intersection non vide avec toutes les autres classes $\mathcal{G}_k$, $k > 1$, et qui reste apprenable par un algorithme à la complexité polynomiale : la classe des GCs *réversibles*. Les grammaires de cette classe n'associent jamais à un même mot de leur vocabulaire deux catégories qui ne se distinguent que par *une seule catégorie de base*. Par exemple, la grammaire $FG(D)$ de l'Exemple 3 est réversible. Dans cet exemple, si on cible une grammaire réversible, il est inutile de procéder à des unifications. Cette notion correspond dans le monde des GCs à celle de *grammaires algébriques réversibles* définie par Sakakibara (Sakakibara, 1992), qui elle-même provenait de celle des automates réversibles d'Angluin (Angluin, 1982). L'algorithme d'apprentissage reproduit le schéma de RG, à la seule différence que l'unification n'est requise qu'entre les catégories affectées à un même mot qui ne diffèrent que par une unique catégorie de base, afin de cibler une grammaire réversible.

L'adaptation des résultats sur les GCs aux grammaires de Lambek a aussi, bien sur, été étudiée. Dans ce domaine, les premiers efforts ont été encourageants : Bonato et Rétoré ont ainsi montré que la classe des grammaires de Lambek rigides (et donc aussi celles des grammaires de Lambek k-valuées, $k \geq 1$) est apprenable par exemples structurés positifs, qui prennent dans ce cas la forme de *structures*

[6]Cette ARC, décrite dans le site http ://www.irisa.fr/paragraphe/Christian.Retore/GRACQ, a été initiée par Christian Rétoré

de preuves (Bonato & Rétoré, 2001). L'algorithme RG, ici encore, s'adapte naturellement. En revanche, l'argument d'une "relation finie" entre une phrase et les structures de preuve possibles correspondantes ne s'applique plus. Et, de fait, aucune classe de grammaires de Lambek k-valuées n'est apprenable à partir de chaînes, comme l'ont montré plus tard Foret et Le Nir (Foret & le Nir, 2002). L'argument décisif, cette fois, est la mise en évidence d'un "point limite" dans la famille des langages produits, qui est une condition suffisante de non-apprenabilité (cf. Figure 4.2). Un tel point limite est une séquence $\langle L_i \rangle_{i \in \mathbb{N}}$ d'éléments de l'ensemble de langages $\mathcal{L}$, telle que $L_0 \subset L_1 \subset ... \subset L_n \subset ...$ et telle que $\cup_{n \in \mathbb{N}} L_n \in \mathcal{L}$.

Dans la perspective que nous avons adoptée, le principal reproche que nous pouvons adresser à ces travaux est leur non prise en compte de la sémantique, d'autant plus dommage que, comme nous l'avons vu, les grammaires catégorielles se prêtent particulièrement bien à la formalisation du Principe de compositionnalité. Le premier à avoir tenté un rapprochement entre l'apprentissage automatique de grammaires catégorielles et leur interprétation sémantique est Adriaans (Adriaans, 1992). Mais lui se plaçait dans une variante du modèle PAC, et il voyait l'acquisition sémantique comme une *conséquence* de l'apprentissage syntaxique, et non comme un préalable facilitateur. Nous avons suivi une autre voie. Tous les éléments sont maintenant en place pour exposer les chemins que nous avons suivis.

Chapitre 5

Contributions personnelles

Les travaux que nous présentons ci-dessous s'étalent sur une dizaine d'années. Ils se consacrent à l'apprentissage de diverses classes de grammaires, à partir de divers types d'exemples. Ils se répartissent naturellement suivant trois grandes approches, correspondant à différentes conceptions du Principe de compositionnalité : la première se concentre sur les *structures de dérivation*, la deuxième sur sa *lexicalisation*. La troisième partie regroupe des recherches où la sémantique n'est pas explicitement disponible, mais qui bénéficient toutefois indirectement de la prise en compte de la compositionnalité. Pour chacune de ces parties, nous avons sélectionné deux articles significatifs reproduits à la fin de ce document.

5.1 La compositionnalité sans dessus-dessous

L'origine de nos travaux est l'idée que, pour modéliser l'apprentissage du langage naturel, il faut partir de données syntaxico-sémantiques, constituées d'énoncés grammaticalement corrects (par exemple "un chat dort") associés à leur représentation sémantique (dans ce cas, la formule : $\exists x[chat'(x) \wedge dort'(x)]$), en faisant l'hypothèse que les secondes dérivent compositionnellement des premières. Le Principe de compositionnalité est habituellement présenté comme une manière de *déduire la sémantique de la syntaxe*. Or, nous proposons plutôt de l'utiliser pour *acquérir la syntaxe à partir de la sémantique*. C'est donc un *renversement de perspective* que nous proposons ici, représenté très schématiquement par la flèche en gras dans le schéma de la Figure 5.1. Ce renversement sera instancié de deux façons différentes, et il donnera lieu finalement à la formulation d'un "nouveau Principe" qui complète le Principe de compositionnalité en l'adaptant au contexte de l'apprentissage.

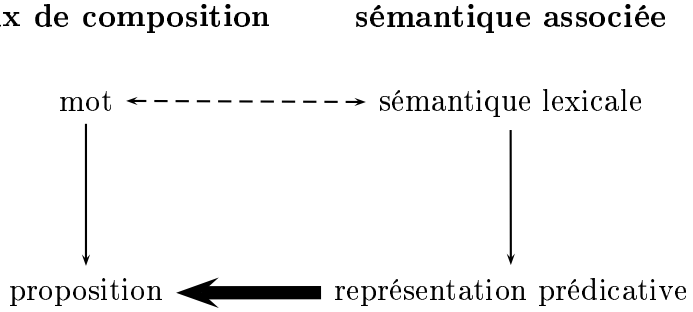

FIG. 5.1 – schéma de l'apprentissage via l'"inversion de la compositionnalité"

5.1.1 Spécialisation contrôlée par la sémantique

Le premier article reproduit dans la suite de ce document a été présenté lors de la conférence ICGI ("International Colloquium on Grammatical Inference") en 98 (Tellier, 1998b). Il reprend, en l'adaptant au système de Lambek, la démarche appliquée aux grammaires catégorielles de type AB dans (Tellier, 1998c; Tellier, 1998a). Cet article est assez programmatique. Son principal intérêt est d'introduire le projet de recherche qui sera poursuivi dans les années à suivre, en affirmant d'emblée dans son titre ("Meaning Helps Learning Syntax") la primauté de la sémantique dans l'acquisition de la syntaxe. Toutefois, l'article n'intègre encore aucun modèle d'apprentissage précis, et aucun résultat formel n'y est donc démontré.

L'algorithme d'apprentissage suggéré (même si sa description est succinte) vaut néanmoins la peine qu'on s'y attarde un peu. Il propose d'associer par défaut une catégorie basique distincte à tout élément du vocabulaire, et de décomposer ces catégories en catégories fractionnaires (du type A/B ou $B\backslash A$ où A et B sont eux-mêmes décomposables) "à la volée" en cherchant à faire une analyse syntaxique de l'énoncé courant. La traduction sémantique est utilisée uniquement *a posteriori* pour sélectionner, parmi les analyses possibles, celles qui se traduisent correctement, via l'isomorphisme de Curry-howard, dans la formule logique globale associée à l'énoncé. Le mécanisme d'inférence employé ici est donc une *spécialisation contrôlée par la sémantique*. En effet, même si ce n'est pas dit explicitement, les catégories affectées doivent être considérées comme des *variables*. L'affectation de base, où une variable distincte est associée à chaque mot du vocabulaire, constitue *la forme générale de toute grammaire rigide* et chaque "décomposition" d'une variable en une fraction de variables est un opérateur de spécialisation (puisque l'introduction d'un opérateur / ou \ rend la catégorie plus spécifique). Le problème posé par les grammaires non rigides est évoqué à la fin de l'article, où une solution empirique est proposée.

L'implémentation et l'étude de cet algorithme a fait l'objet du stage de DEA de Julien Gest en 1998 (Gest, 1998). A l'époque, la non apprenabilité des grammaires de Lambek rigides à partir de chaînes (Foret & le Nir, 2002) n'était pas connu. De fait, pour garantir que l'algorithme s'arrête, Gest montre que l'on doit élaguer les analyses syntaxiques possibles. L'étude de la restriction en termes de langage de cet élagage n'a pas été menée plus avant. Quelques tests expérimentaux ont pu être faits à l'époque, à une échelle encore modeste : l'algorithme est clairement exponentiel et il n'est pas raisonnable d'envisager de l'employer sur des corpus réels.

Rappelons que l'apprentissage par spécialisation semble cognitivement plus réaliste que par généralisation (cf section 2.2.3). La technique décrite dans cet article a été reprise plus précisément pour les grammaires AB par Moreau dans (Moreau, 2004), mais sans faire appel à la sémantique pour faire un choix parmi les grammaires candidates. Pour que l'algorithme converge au sens de Gold, il faut pourtant donner un résultat unique, ce qui impose en théorie de faire des tests d'inclusion entre langages (en bornant la taille des exemples pour que cela reste décidable, cf. partie 4.2.3).

De son côté, Fulop a récemment publié des résultats d'apprentissage de grammaires logiques qui exploitent l'isomorphisme de Curry-Howard, mais en adaptant les techniques de généralisation employées par Kanazawa (Fulop, 2005).

5.1.2 Le sens des structures

Le deuxième article, présenté en 99 au Amsterdam Colloquium (Tellier, 1999c), adopte une approche différente. L'objectif cette fois est de se rattacher aux résultats de Kanazawa sur l'apprentissage de la classe des grammaires catégorielles de type AB k-valuées à partir d'exemples structurés. Plutôt que de chercher un nouvel algorithme d'apprentissage, nous proposons cette fois de *justifier par la sémantique l'usage des exemples structurés*. Notre proposition se situe donc *en amont* de l'apprentissage proprement dit ; il vise à expliquer d'où proviennent les données d'entrée dont Kanazawa a besoin, comme le montre la Figure 5.2

Le point de départ de cet article est le suivant : la donnée de la sémantique lexicale associée aux mots présents dans un énoncé et de l'exemple structuré correspondant, qui constitue en quelque sorte le *squelette* de son analyse syntaxique, suffisent pour reconstituer la sémantique globale de cet énoncé. En effet, la structure donne le parenthésage de l'énoncé et l'étiquette *FA* ou *BA* de chaque noeud donne le sens des applications fonctionnelles à appliquer sur les formules associées aux mots du vocabulaire. Par exemple, à partir de l'exemple structuré : *FA* (*FA* (un chat) dort) et des traductions logiques de "un" : $\lambda P \lambda Q \exists x[P(x) \wedge P(x)]$, "chat" : $chat'$ et "dort" : $dort'$, on retrouve facilement : $\exists x[chat'(x) \wedge dort'(x)]$ (cf. Figure 3.2).

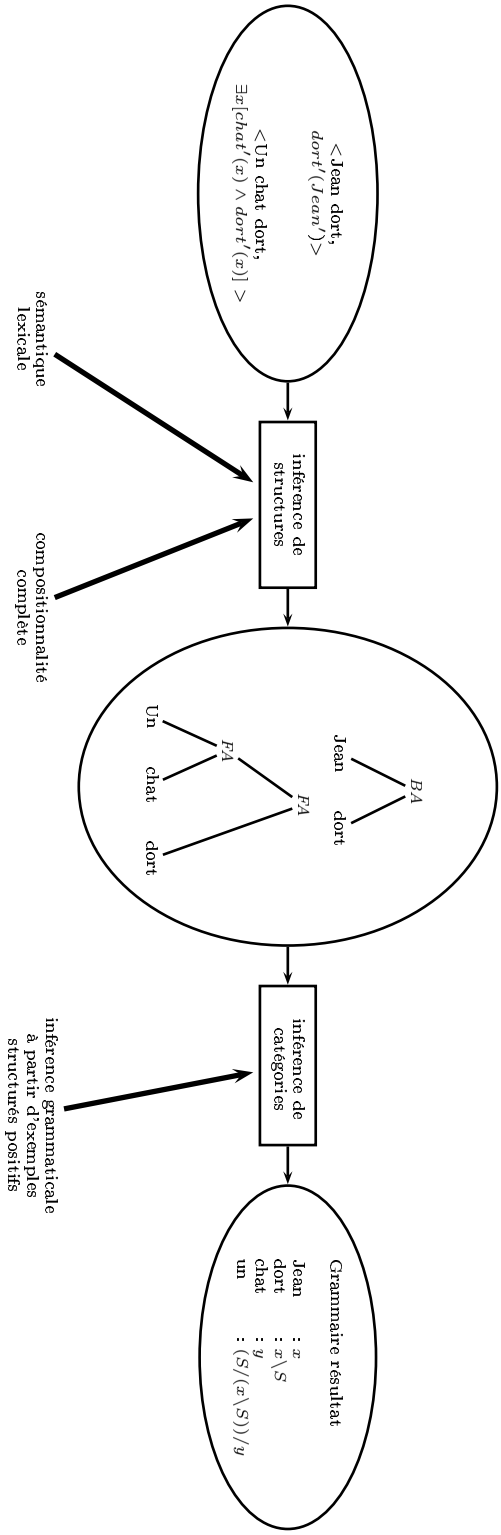

FIG. 5.2 – apprentissage syntaxico-sémantique en deux étapes

Mais on a vu que dans une modélisation "naturelle" de l'acquisition syntaxique, ce sont plutôt les traductions logiques associées aux mots et à la phrase qui sont disponibles, alors que les exemples structurés, eux, sont inconnus. Néanmoins, dans le cas des GCs, le nombre de FA-structures pouvant être associées à un énoncé donné est fini et il est facile de les énumérer (c'est d'ailleurs l'argument qui a permis à Kanazawa de garantir l'apprentissage des GCs k-valuées à partir de chaînes). Parmi celles-ci, figure toujours au moins un exemple structuré. Ainsi, si on part de la phrase "un chat dort" et si on cherche quel exemple structuré lui est associé, on est amené à envisager tous les arbres binaires possibles, avec toutes les étiquettes FA et BA possibles à leurs noeuds, dont cette phrase est le feuillage. Parmi ceux-ci, en intégrant maintenant les traductions logiques aux feuilles, seul celui de la Figure 4.3 (à droite) donne lieu à la bonne formule logique finale. Les traductions sémantiques permettent d'évaluer la qualité d'une FA-structure candidate au statut d'exemple structuré.

Cela nous amène à exiger une notion de la compositionnalité un peu plus forte que dans sa formulation standard à base de simple "morphisme" : il faut en effet non seulement que les exemples structurés se traduisent en arbres sémantiques corrects, mais aussi que *toute FA-structure donnant lieu à une traduction sémantique correcte soit nécessairement un exemple structuré* (cf. l'article pour une formulation plus rigoureuse). Cette nouvelle formulation traduit l'intuition que *la structure est porteuse de sens*. Nous appelons "compositionnalité complète" cette propriété. Si elle est assurée, on dispose d'un mécanisme de type générer-tester pour fournir des exemples structurés à un algorithme d'apprentissage.

Cette idée a aussi été reprise et développée dans (Tellier, 1999b; Tellier, 1999a) et a été le sujet du stage de DEA de Daniela Dudau (Dudau, 2000), encadrée avec Marc Tommasi. Cette fois aussi, un programme a été implémenté et testé sur des exemples jouets.

5.1.3 Vers un nouveau Principe

Les travaux évoqués précédemment nous ont donc amené progressivement à reconsidérer le Principe de compositionnalité sous un angle nouveau. Il y a encore un écart important entre la caractérisation du Principe de compositionnalité par l'*isomorphisme* de Curry-Howard, qui accorde une place symétrique à la syntaxe et à la sémantique, et son usage usuel qui fait dériver la seconde de la première.

Il est au moins un domaine où il nous semble naturel d'*inverser radicalement* le sens du Principe de compositionnalité, pour faire dériver la syntaxe de la sémantique : celui de l'*émergence du langage*. Nous estimons en effet, comme Bickerton (Bickerton, 1990) (cf. partie 2.2.1), que les langues naissent de la faculté de représentation sémantique. La syntaxe ne ferait alors que traduire sur le plan des signifiants les *combinaisons de sens* qui s'opèrent au niveau des signifiés, pour

rendre compte de faits de plus en plus élaborés.

Reprenons l'exemple des traductions logiques des mots "un", "chat" et "dort". Elles reflètent le fait que "un" est porteur d'une structure propositionnelle qui doit être instanciée par la donnée de deux prédicats à un argument, rôles joués par les deux autres traductions. Mais ces formules, à elles seules, *n'imposent aucun ordre des mots dans la phrase*. A chaque fois qu'on fixe un tel ordre, on fixe par la même occasion la FA-structure permettant l'exécution des applications fonctionnelles sur ces traductions, et donc indirectement la syntaxe sous-jacente.

Toutefois, pour qu'il existe effectivement une grammaire k-valuée compatible avec un ensemble de FA-structures, celles-ci doivent présenter des régularités. La valeur de k est en quelque sorte inversement corrélée avec le degré de régularité requis : si $k = 1$, chaque mot ne peut jouer qu'un seul rôle syntaxique et leur ordre est contraint (par exemple, les verbes intransitifs ne peuvent recevoir qu'une seule catégorie correspondant soit à S/T soit à $T\backslash S$, qui fixe leur position vis-à-vis de leur sujet). Plus k est grand, plus on s'accorde de la souplesse mais plus l'apprentissage risque d'être difficile. C'est sans doute un compromis tendant à *minimiser ce paramètre k tout en préservant de l'expressivité* qui, à notre sens, régit l'émergence des langues naturelles.

C'est cette idée que nous avons esquissée dans (Tellier, 2000), présentée en poster à la conférence "Evolution of Language". Cet article constitue en quelque sorte une version radicale et vulgarisée de celui cité dans la section précédente. Il aboutit à la proposition d'un "scénario" rudimentaire d'émergence, qui reproduit les étapes du schéma de la Figure 5.2, à la différence près que la première étape n'est composée que de combinaisons sémantiques. Ici encore, les exemples structurés jouent le rôle d'étape intermédiaire sur le chemin de la spécification d'une grammaire complète.

Le domaine de l'apprentissage d'une langue existante est très similaire, sauf que les énoncés entendus contraignent un peu plus le scénario. En fait, le vrai "Principe" sous-jacent mis en oeuvre dans nos recherches peut s'énoncer de la manière suivante : *une grammaire est complètement spécifiée par le sens de ses unités, la manière dont ces sens sont compositionnellement combinés pour former une proposition et la succession linéaire des unités qui la signifient.* Ou encore, en version plus synthétique : *la syntaxe ne dépend que de la sémantique compositionnelle qu'on peut lui associer, et de l'ordre des mots*[1]. Cette "compositionnalité inversée" (qui intègre dans sa formulation la compositionnalité au sens traditionnel) est le fondement de tous les travaux évoqués jusqu'ici, ainsi que de tous ceux de la sous-partie suivante. Elle justifie l"'apprenabilité" des langues naturelles. Elle explique que pour "apprendre" une langue inconnue, il faut disposer à la fois d'énoncés de

[1] cette formulation vaut surtout pour les langues à ordre fixe, les seules correctement modélisées par les GCs ; pour les autres, ce sont les "cas" qui doivent se substituer à l'ordre des mots

cette langue et de leur sens. Elle permet ainsi peut-être de mieux comprendre pourquoi la pierre de Rosette était nécessaire au déchiffrement des hiéroglyphes égyptiens...

5.2 Apprentissage à partir d'exemples typés

Ces premiers travaux, très généraux, ne permettaient pas encore d'envisager des implémentations conséquentes. De plus, ils n'exploitaient pas vraiment une propriété pourtant fondamentale des grammaires catégorielles : leur caractère *lexicalisé*. C'est en cherchant à tirer profit de cette propriété que nous avons été amenée à définir une nouvelle situation d'apprentissage, où des informations sémantiques lexicalisées sont fournies à l'apprenant sous la forme de *types logiques dérivés des catégories syntaxiques*. Cette nouvelle approche est schématiquement visualisée par la flèche en gras sur la Figure 5.3. Nous l'avons d'abord étudiée dans le cadre des GCs, avant de l'étendre à d'autres classes de grammaires.

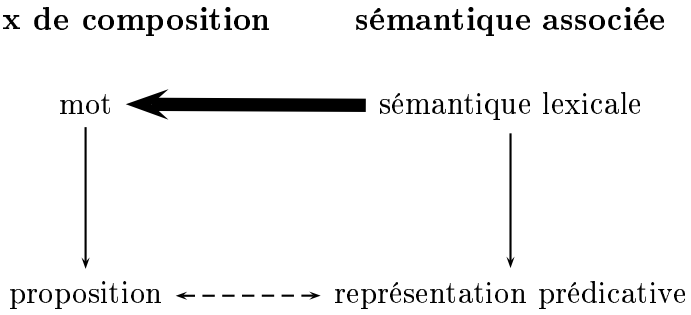

FIG. 5.3 – schéma de l'apprentissage "à partir d'exemples typés"

5.2.1 Apprenabilité des grammaires AB à partir d'exemples typés

Rappelons que les représentations sémantiques que nous employons sont issues d'une *logique typée*. Le typage utilisé, inspiré de celui de Montague, est très standard dans le domaine. Nous avons vu en partie 3.3.2, particulièrement dans l'Exemple 5.2.1, que la compositionnalité induit un morphisme h qui transforme les catégories syntaxiques dans ces types logiques. Dans la perspective de la primauté de la sémantique, il est naturel de se demander si, partant des types, il est possible de "remonter" aux catégories. L'étude cette possibilité a fait l'objet de la thèse de Daniela Dudau-Sofronie (Sofronie, 2004), co-encadrée avec Marc Tommasi.

L'hypothèse fondatrice de ce travail est donc que les données disponibles à l'apprenant sont des *exemples typés*, c'est-à-dire des énoncés syntaxiquement corrects où chaque mot est associé à un type : "$(un, \langle\langle e,t\rangle, \langle\langle e,t\rangle, t\rangle\rangle)(chat, \langle e,t\rangle)(dort, \langle e,t\rangle)$" est l'exemple typé correspondant à notre exemple prototypique. Le vocabulaire sur lequel sont construits ces exemples typés est donc $\Sigma \times \mathcal{T}$.

Remarquons tout d'abord que l'apprenabilité par exemples positifs seuls des GCs rigides implique l'apprenabilité par exemples positifs seuls des GCs G dont chaque affectation distincte de catégorie de la forme $\langle v, C\rangle \in G \subset \Sigma \times Cat(\mathcal{B})$ donne lieu à une unique association $\langle v, \tau\rangle \in \Sigma \times \mathcal{T}$ où $\tau = h(C)$. Nous appelons $\mathcal{G}_h$ cette classe de GCs. En effet, sur le nouveau vocabulaire $\Sigma \times \mathcal{T}$ considéré, ces grammaires sont rigides. En fait, pour un h fixé, il est même facile de montrer qu'il existe un nombre fini de GCs de $\mathcal{G}_h$ (à un renommage près et sans catégories inutiles) vérifiant cette propriété et compatibles avec tout ensemble d'exemples typés : l'argument de la "cardinalité finie" (qui figure dans le schéma de la Figure 4.2) confirme l'apprenabilité de la classe $\mathcal{G}_h$.

Nous faisons figurer dans ce document un article de synthèse présenté à CAP, la conférence française d'apprentissage automatique, en 2003 (Dudau-Sofronie et al., 2003b) couvrant les principaux résultats obtenus sur cette base. Les publications intermédiaires se sont focalisées soit sur l'algorithme d'apprentissage proposé (Dudau-Sofronie et al., 2001a; Dudau-Sofronie et al., 2001b), soit sur les conditions de validité du résultat d'apprenabilité de la nouvelle classe de GCs ainsi caractérisée, et sur ses propriétés théoriques (Dudau-Sofronie et al., 2003a). Nous avons ainsi montré que pour tout langage engendré par une GC, il existe un morphisme h et une grammaire de $\mathcal{G}_h$ qui génère le même langage de structures et qui est apprenable par exemples typés. La notion de "rigidité étendue" que nous exploitons constitue donc un gain significatif d'expressivité, au prix de données d'entrée plus riches. Nous ne ferons ici que reprendre les caractéristiques principales de l'algorithme d'apprentissage proposé, pour mieux le comparer à ceux déjà évoqués.

Ce qui distingue une catégorie syntaxique (au sens des GCs) d'un type logique, c'est d'une part l'ensemble de ses éléments atomiques ($\mathcal{B}$ dans le premier cas, en général $\{e,t\}$ dans le second) et d'autre part la présence des opérateurs / et \ au niveau syntaxique, qui disparaissent au niveau sémantique. Mais les catégories et les types peuvent tous les deux être vus comme des *termes* reliés par le morphisme h. Pour reconstituer les catégories[2] à partir des types, nous proposons tout d'abord de "variabiliser" les exemples typés, en introduisant des variables aux positions où les opérateurs / et \ devraient figurer. Ces variables sont distinctes chaque fois qu'un couple $\langle mot, type\rangle$ distinct est pré-

[2]dans cette partie où la nature de termes des catégories est cruciale, on emploie la notation qui utilise les opérateurs comme *préfixes*

sent dans les exemples typés. L'exemple typé précédemment cité devient ainsi : $(un, x_1\langle x_2\langle e,t\rangle, x_3\langle x_4\langle e,t\rangle, t\rangle\rangle)(chat, x_5\langle e,t\rangle)(dort, x_6\langle e,t\rangle)$.

La technique d'apprentissage proposée consiste alors à essayer d'effectuer une *analyse syntaxique* des exemples typés variabilisés, donnant comme résultat $t = h(S)$, en déduisant au fur et à mesure des *contraintes* sur la valeur des variables x_i qu'ils contiennent, à la manière des algorithmes de spécialisation évoqués en partie 5.1.1. Ces contraintes peuvent prendre deux formes distinctes : soit $x_i = /$ ou $x_i = \backslash$, pour un indice i, soit $x_i = x_j$ pour deux indices $i \neq j$. Pour notre exemple canonique élémentaire, une seule analyse syntaxique est possible et elle mène à la *conjonction* des contraintes suivantes : $x_1 = /$, $x_2 = x_5$, $x_3 = /$ et $x_4 = x_6$. Dans les cas où plusieurs analyses sont possibles, on obtient une *disjonction de conjonctions de contraintes.*

Les contraintes du type $x_i = /$ ou $x_i = \backslash$ correspondent bien à une opération de *spécialisation.* Les contraintes d'égalité entre variables peuvent, aussi, mener à une telle spécialisation : par exemple, quand des contraintes de la forme $x_i = x_j$ et $x_i = /$ entraînent que $x_j = /$. Mais ce n'est pas toujours le cas. Pour le comprendre, imaginons qu'en plus du premier exemple typé, on en dispose aussi d'un deuxième, variabilisé comme suit : $(un, x_1\langle x_2\langle e,t\rangle, x_3\langle x_4\langle e,t\rangle, t\rangle\rangle)(homme, x_7\langle e,t\rangle)(court, x_8\langle e,t\rangle)$. Les contraintes obtenues à partir de ce deuxième exemple, parallèles à celles du premier, sont : $x_1 = /$, $x_2 = x_7$, $x_3 = /$ et $x_4 = x_8$. Il reste à caractériser la ou les grammaire(s) de $\mathcal{G}_h$ compatible(s) avec la conjonction de ces contraintes : pour cela, nous définissons l'ensemble des catégories de base comme *l'ensemble des classes d'équivalence des sous-types variabilisés maximaux dont aucune variable interne x_i n'est concernée par une contrainte de la forme $x_i = /$ ou $x_i = \backslash$ et qui sont substituables les uns aux autres par égalité*[3]. Dans notre exemple, nous avons deux classes de ce genre : $x_2\langle e,t\rangle = x_5\langle e,t\rangle = x_7\langle e,t\rangle$ et $x_3\langle e,t\rangle = x_6\langle e,t\rangle = x_8\langle e,t\rangle$. Chacune correspond donc à une catégorie de base distincte : la première est la classe des noms communs, la deuxième celle des verbes intransitifs. La présence du contexte commun constitué par le même mot typé $(un, \langle\langle e,t\rangle, \langle\langle e,t\rangle, t\rangle\rangle)$, donc variabilisé avec les mêmes variables, suffit donc à *unifier* les catégories des mots avec lesquels il s'associe. Cette opération conduit, cette fois, à des *généralisations.* Notons bien que les deux classes qui ont été identifiées sont distinctes bien qu'elles sont associées par h au *même type logique.*

La particularité de la technique d'apprentissage mise en oeuvre dans ces travaux est donc qu'elle combine deux démarches : l'ordre des mots, fixé par les opérateurs / et \, est acquis par *spécialisation* au fur et à mesure de l'analyse des exemples, alors que la définition des catégories syntaxiques de base est induite par *généralisation finale.* Ce compromis est cognitivement assez crédible, tout en évitant l'explosion combinatoire dont souffrent les techniques de spécialisation pure. Pour faire de

[3] en outre, on a toujours : $h(S) = t$ pour les t figurant aux racines des analyses syntaxiques

cette stratégie, dont on a pu prouver la validité et la complétude (Sofronie, 2004), un algorithme d'apprentissage au sens de Gold, il faudrait toutefois être capable de choisir *une* grammaire unique parmi celles induites des contraintes. Pour éviter la surgénéralisation inhérente aux algorithmes qui travaillent par exemples positifs seuls, c'est une de celles produisant le plus petit langage typé possible qui devrait être choisie. Mais nous ne savons même pas si l'inclusion entre langages typés est décidable, et l'astuce de Kanazawa consistant à tester des inclusions sur des énoncés de tailles bornées est inapplicable en pratique.

Nous allons voir maintenant comment nous avons pu mener des tests expérimentaux, et comment étendre cette première approche à d'autre formalismes.

5.2.2 Expérimentations : apprendre le langage des schtroumpfs

L'algorithme présenté dans la section précédente fournit, à partir d'un ensemble d'exemples typés provenant d'une GC de $\mathcal{G}_h$, l'ensemble des GCs de cette classe (à un renommage près et sans catégories inutiles) qui produisent ces exemples typés. Le problème est que le nombre de ces grammaires peut être très grand : il dépend, non pas de la taille des exemples (c'est-à-dire du nombre de mots dans les énoncés), mais du nombre de variables introduites, qui lui-même dépend du nombre de couples $\langle mot, type \rangle$ distincts et de la taille des types présents dans les exemples (cf. exemple dans (Dudau-Sofronie et al., 2003b)).

Malgré tout, cet algorithme est le premier de ceux que nous avons contribué à élaborer dans ce domaine, qui a pu être implémenté dans une plate-forme robuste (Dudau-Sofronie et al., 2002), et testé sur des données réelles. En partie 5.1, en effet, nous nous heurtions à la trop grande complexité des algorithmes décrits et à la difficulté de disposer des couples "énoncé/traduction logique" qu'ils requièrent en entrée. Les types logiques ne sont pas non plus, *a priori*, présents dans les corpus usuels. Néanmoins, ils peuvent plus facilement être *déduits* à partir des étiquettes fournies par les analyseurs lexicaux, en suivant la démarche décrite dans la Figure 5.4. Ce schéma montre que nous avons fait appel à deux ressources externes : le logiciel libre "Tree Tagger", qui associe une étiquette lexicale (du type "nom-commun", "adjectif, "verbe", etc.) à chaque mot d'un texte, et une "table de correspondance", définie à la main, qui associe un ou plusieurs types à chacune de ces étiquettes. La phase de "nettoyage" (III sur le schéma) s'occupe notamment de désambiguiser les affectations de types possibles.

Un corpus de contes écrits par des enfants à été recueilli et soumis à ce traitement (ce qui a fait l'objet d'un stage de maîtrise de science du langage de Thomas Desvenain), avant de servir de données d'entrée à notre algorithme. Les expériences ont été réalisées par Daniela Dudau-Sofronie. L'efficience du programme a été mesurée par le *nombre de grammaires résultats* obtenues en sortie. Nous avons pu constater que deux facteurs influencent ce nombre : l'ordre de présentation des

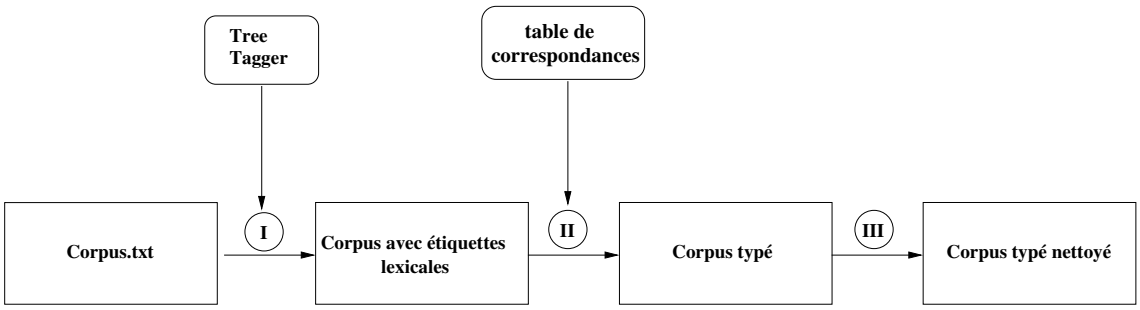

FIG. 5.4 – traitements automatiques successifs pour obtenir un corpus typé

exemples (on a intérêt à présenter les *phrases courtes d'abord*) et la redondance du vocabulaire (on a intérêt à répéter l'usage du plus de mots possibles). Avec notre corpus, en tirant aléatoirement 15 phrases comprenant en tout 85 mots dont 81 différents, on obtient plus de 2500 grammaires différentes (moyenne sur 8 tirages).

Pour aller plus loin, il faut bien sûr réduire ce nombre. Pour cela, nous avons forcé les répétitions de vocabulaire, en opérant des *substitutions* dans le corpus : pour certaines étiquettes lexicales (les noms communs, les pronoms, les déterminants et les verbes transitifs), le nombre de mots recevant cette étiquette a été réduit à deux instances (tirées au sort). C'est un peu comme si on traduisait les textes initiaux dans un "langage des schtroumpfs" disposant de peu de mots distincts (sans empêcher d'ailleurs que le même mot "schtroumpf" puisse aussi bien être un nom qu'un verbe !). Les résultats confirment l'influence de l'ordre des exemples, mais sont encore limités quant à l'évaluation de la pertinence linguistique des grammaires résultats (Sofronie, 2004; Dudau-Sofronie & Tellier, 2004b).

5.2.3 Extension à d'autres classes de grammaires

Nous avons aussi, bien sûr, cherché à étendre ce modèle d'apprentissage à d'autres classes de grammaires que les GCs. Le candidat suivant le plus naturel est constitué des grammaires de Lambek. Les classes de grammaires de Lambek k-valuées ne sont en effet pas apprenables à la limite à partir de chaînes, mais le sont à partir d'exemples structurés d'une certaine forme normale. Le sont-elles à partir d'exemples typés ? La réponse est oui, à condition de se restreindre à la classe des grammaires de Lambek qui correspond à la définition de la classe $\mathcal{G}_h$ pour les grammaires de type AB. L'argument qui le prouve est en fait très simple : les catégories grammaticales étant toujours plus "petites" (au sens du nombre d'instances de catégories de base qu'elles contiennent) que les types qui leur correspondent par un morphisme, il existe un nombre fini de catégories compatibles avec un type donné. L'argument de la "cardinalité finie" s'applique donc au niveau des grammaires de

Lambek de notre classe produisant un échantillon d'exemples typés.

L'algorithme esquissé en partie 5.2.1, consistant à introduire des variables dans les types et à chercher les ensembles de contraintes qu'elles doivent vérifier pour qu'une analyse syntaxique soit possible, est reproductible pour les grammaires de Lambek. Cette stratégie a été le sujet du stage de DEA de Fredéric Dupont, qui a implémenté la méthode (Dupont, 2003). Les résultats théoriques ont, eux, été présentés à la conférence "Categorial Grammars" (Dudau-Sofronie & Tellier, 2004a).

Ces travaux nous suggèrent que les types contraignent fortement les structures syntaxiques possibles. En fait, les types ne sont rien d'autre eux-mêmes que des *structures lexicalisées*, et le morphisme h qui transforme les catégories grammaticales en types réalise la *lexicalisation du Principe de compositionnalité*.

Une dernière mise en oeuvre assez différente de ces mêmes idées a été étudiée dans le cadre des *grammaires de prégroupe*, un nouveau formalisme récemment introduit par... Lambek ! (Lambek, 2001; Lambek & Casadio, 2002). Ce modèle algébrique lexicalisé permet de rendre compte de façon élégante de phénomènes linguistiques subtils, comme l'ordre des déictiques en français et en italien (Bargelli & Lambek, 2001; Lambek & Casadio, 2001). Il n'a malheureusement pas pu encore être rigoureusement mis en correspondance avec une interprétation sémantique, malgré quelques tentatives préliminaires (Preller, 2005). Les grammaires de prégroupes ont l'expressivité des langages algébriques (Buszkowski, 2001), et ne sont pas apprenables à la limite à partir de chaînes (Béchet & Foret, 2003).

Nous avons donc cherché une classe de grammaires de prégroupe qui soit apprenable à partir de chaînes enrichies par des informations lexicalisées, comme notre classe de grammaires de Lambek l'est à partir d'exemples typés. Cela nous a amené à généraliser l'approche fondée sur les types logiques, et la technique d'apprentissage par spécialisation donnant lieu à des ensembles de contraintes, exposé dans cette section. Ce travail est décrit dans (Béchet et al., 2004), présenté à ICGI en 2004 et reproduit dans ce document.

Par rapport aux travaux précédents, cet article présente des avancées intéressantes. Tout d'abord, l'argument d'apprenabilité de la classe ne relève plus de la "cardinalité finie". Mais c'est surtout au niveau de l'algorithme d'apprentissage que les choses changent. Le domaine des variables introduites ne se réduit plus à l'ensemble des opérateurs $\{/, \backslash\}$, comme c'était le cas précédemment : il couvre l'ensemble des entiers relatifs. Du coup, les contraintes produites par la recherche d'analyses syntaxiques sont également plus complexes. Surtout, nous montrons que ces contraintes spécifient dans certains cas un nombre de grammaires exponentiellement plus grand que leur taille ; c'est ce critère qui justifie en général l'introduction de contraintes, et encourage à les résoudre le plus tard possible. Ainsi, la stratégie globale de recherche des grammaires compatibles avec un ensemble d'exemples

enrichis se trouve découpée en deux étapes successives indépendantes : une étape de *production de contraintes* et une étape de *résolution de contraintes*. Il faudrait, comme précédemment, faire suivre ces étapes d'une phase de tests d'inclusions entre langages enrichis, pour rester dans le cadre du modèle de Gold. La validité théorique de cette approche a été démontrée (article en soumission), mais n'a pu encore être confirmée par une implémentation. Il reste que les données d'entrée nécessaires au fonctionnement de la méthode ne sont pas faciles à justifier, d'autant que leur lien avec la sémantique n'est pas complètement établi.

5.3 Apprentissage sans sémantique

On a beau être convaincu de la primauté de la sémantique sur la syntaxe, il faut bien reconnaître que l'information sémantique n'est en général pas explicitement disponible dans les données textuelles, qui sont la matière première de la plupart des travaux d'ingénierie linguistique. Et en inférence grammaticale "classique", on ne suppose pas non plus disposer de cette information.

Mais il se trouve que les détours que nous avons pris pour les travaux présentés jusqu'à présent nous ont aussi suggéré de nouvelles pistes pour l'apprentissage "sans sémantique". Nous regroupons sous ce label des travaux assez différents. Les premiers, dans la lignée directe des recherches présentées jusqu'ici, sont consacrés à l'inférence de GCs à partir de *chaînes de mots*, plus efficacement que le faisait Kanazawa. Le dernier relève de domaines apparemment assez différents de l'environnement envisagé dans l'ensemble de ce document (il y est question d'extraction d'information et de systèmes Question/Réponse) mais on peut aussi le voir comme un problème d'apprentissage de grammaires locales à partir de textes étiquetés.

Ces démarches sont grossièrement représentées par la flèche en gras dans le schéma de la Figure 5.5. La dimension sémantique y figure encore, mais au seul titre de "fantôme" qui n'influence qu'indirectement les choix faits. Cette dernière partie pourrait aussi s'intituler : de l'intérêt du Principe de compositionnalité comme inconscient linguistique...

5.3.1 Langages réguliers et GCs

Les GCs, on l'a vu, sont capables d'engendrer tous les langages algébriques. Et, pour tout entier $k \geq 1$, la classe $\mathcal{G}_k$ est apprenable à la limite à partir de chaînes.

Mais l'apprentissage au sens de Gold de classes de grammaires à partir de chaînes a surtout été étudié dans un tout autre contexte que celui des GCs : celui des langages réguliers, représentés par des automates à états finis. Nous avons évoqué certains de ces travaux au fil de ce document ; nous y revenons brièvement dans cette section, un peu plus précisément dans la section suivante. Puisque les

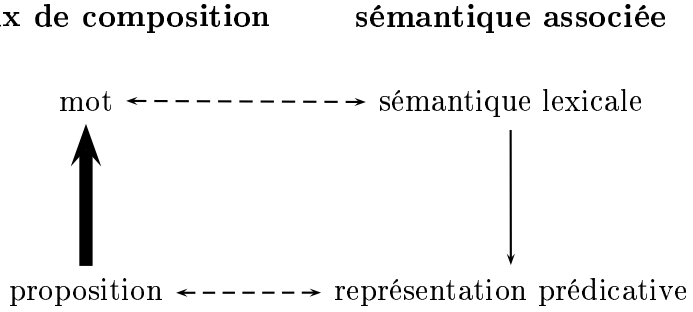

FIG. 5.5 – schéma de l'apprentissage "sans sémantique"

langages algébriques recouvrent les langages réguliers, nous nous sommes tout naturellement demandé si les résultats obtenus de part et d'autre se recoupaient.

Pour établir des rapprochements, nous avons cherché à caractériser une sous-classe des GCs qui engendre l'ensemble des langages réguliers. Nous avons ainsi introduit $\mathcal{G}_r$, classe des GCs dont toutes les affectations de catégories sont de la forme $\langle v, A\rangle$ ou $\langle v, A/B\rangle$, pour $v \in \Sigma$ et $A, B \in \mathcal{B}$ (A et B sont donc des catégories de base). Comme attendu $L(\mathcal{G}_r) = \{L(G)|G \in \mathcal{G}_r\}$ coïncide avec l'ensemble des langages réguliers. Et les résultats de Kanazawa impliquent que pour tout $k \geq 1$, $\mathcal{G}_k \cap \mathcal{G}_r$ est apprenable par exemples positifs seuls. La classe des "strictly deterministic automata", proposée par (Yokomori, 1995) se révèle d'ailleurs coincider avec $\mathcal{G}_1 \cap \mathcal{G}_r$.

De plus, les exemples structurés produits par les grammaires de $\mathcal{G}_r$ sont exclusivement des "peignes droits" (c'est-à-dire des arbres binaires dont tous les fils gauches sont des feuilles) et tous leurs noeuds internes sont étiquetés par *FA*. Pour les grammaires de $\mathcal{G}_r$, les exemples structurés se déduisent donc linéairement des chaînes.

Que se passe-t-il quand on applique l'algorithme d'apprentissage par exemples structurés de Kanazawa à un échantillon composés uniquement de tels peignes ? Le mieux est de l'illustrer sur un exemple...

Exemple 4. *Soit $\Sigma = \{a, b\}$ et D l'ensemble constitué des peignes droits avec noeuds internes FA associés aux chaînes "ab" et "aabb". On applique l'étape* 1 *de l'algorithme RG (cf. partie 4.2.2) à cet ensemble : on obtient les arbres de la figure 5.6.*

La forme générale $FG(D)$ correspondante est alors définie comme suit :

- *a : $S/x_1,\ S/x_4,\ x_4/x_3$;*
- *b : $x_1,\ x_3/x_2,\ x_2$.*

La suite de l'algorithme d'apprentissage consiste à chercher des substitutions unificatrices *applicables sur $FG(D)$. Soit par exemple σ telle que : $\sigma(x_4) = \sigma(x_1) =$*

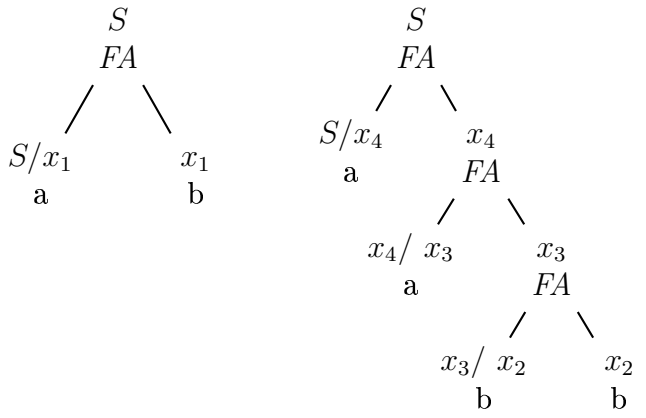

FIG. 5.6 – résultat de la première étape de l'algorithme RG à $D = \{ab, aabb\}$

x_3/x_2 (et σ est l'identité partout ailleurs). La GC $\sigma(GF(D))$ est alors définie par :

- *$a : S/(x_3/x_2),\ (x_3/x_2)/x_3$;*
- *$b : x_3/x_2,\ x_2$.*

Cette GC est 2-valuée mais elle n'appartient plus à la classe $\mathcal{G}_r$: en fait, elle engendre le langage $a^n b^n$! Pour s'en convaincre, on montre en Figure 20 l'analyse syntaxique de "aaabbb". Cette analyse est rendue possible par le fait que σ a unifié l'étiquette d'un noeud interne (x_4) avec celle d'une feuille (x_3/x_2) ouvrant ainsi la possibilité de substituer le sous-arbre de racine x_4 à la place de la feuille étiquetée par x_3/x_2. La structure résultante n'est plus un peigne mais un peigne de peignes.

L'Exemple 4 montre qu'à partir d'exemples qui sont des peignes, on peut inférer une grammaire donc l'expressivité dépasse la classe des langages réguliers, en produisant plus que des peignes. Cela nous a amenée à définir une nouvelle sous-classe de GCs : la classe $\mathcal{G}_k^{FA} = \{\sigma(G) | G \in \mathcal{G}_k \cap \mathcal{G}_r$ et G est sans catégorie inutile et σ est une substitution unificatrice pour $G\}$. Les éléments de cette classe produisent des "peignes de peignes" et sont apprenables à partir d'un ensemble caractéristique constitué uniquement de peignes. Pour les apprendre à partir de chaînes, on peut donc éviter de tester toutes les structures sous-jacentes possibles, comme le suggérait Kanazawa : leur associer des peignes avec noeuds *FA* suffit, à condition d'adapter un peu l'algorithme d'apprentissage. C'est ce que nous avons détaillé dans (Tellier, 2005c; Tellier, 2005b). Le premier de ces articles, présenté à la conférence "Logical Aspects of Computational Linguistics" en 2005, est reproduit dans ce document.

Bien sûr, pour tout $k \geq 1$, $\mathcal{G}_k^{FA} \subset \mathcal{G}_k$. De plus, $\bigcup_{k\geq 1}\{L(G) | G \in \mathcal{G}_k^{FA}\}$ contient tous les langages régulier mais, malheureusement, pas tous les langages algébriques : $a^n b^n c$, par exemple, ne peut être engendré par des peignes de peignes

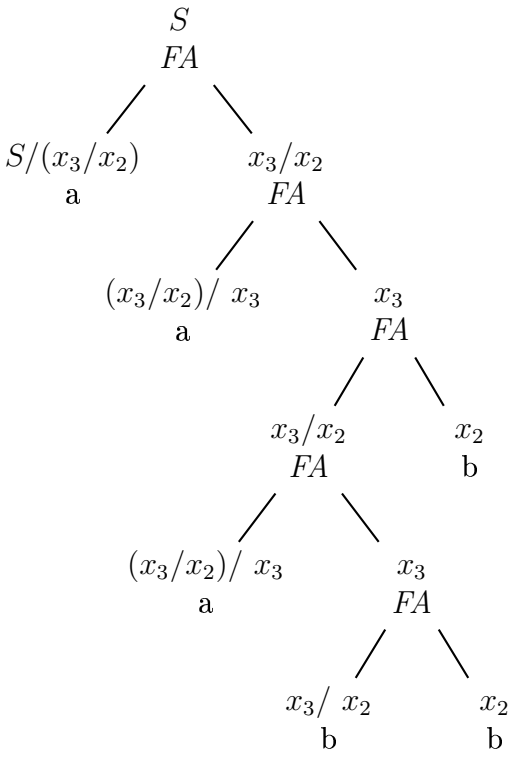

FIG. 5.7 – arbre d'analyse syntaxique de *aaabbb* par $\sigma(FG(D))$

droits (on ne peut pas à la fois produire la duplication des "a" et des "b" tout en empêchant celle de "c"). Mais nous ne sommes pas à l'heure actuelle capable de caractériser plus précisément l'extension de cette classe.

Apprendre des grammaires algébriques à partir de chaînes est fondamentalement difficile, parce que ces grammaires produisent des *structures arborescentes qui sont sous-déterminées par la donnée de leur feuillage*, et qu'*il existe un saut exponentiel entre une chaîne et l'ensemble des structures possibles qu'on peut lui associer*. Dans les sections précédentes, nous avions vu que la sémantique donnait un moyen de sélectionner les structures. A défaut de sémantique, toutes les structures se valent ! Dans ce cas, notre travail montre qu'un bon moyen de procéder pour éviter l'explosion combinatoire est de fixer une forme normale pour les structures qui soit la moins coûteuse possible à obtenir à partir des chaînes : la notion de "peigne de peignes" a alors toute sa pertinence.

5.3.2 L'art de raccorder les treillis

Les résultats évoqués dans la section précédente peuvent aussi se voir suivant un angle différent, qui privilégie la structure de l'espace de recherche. Nous avons vu en partie 4.2 que l'ensemble des GCs rigides, muni de l'opération de substitution, constitue un *treillis*. Pour les classes $\mathcal{G}_k$ avec $k > 1$, on perd la propriété de l'existence d'un unique plus petit unifieur pour tout ensemble de grammaires, mais l'application d'une substitution quelconque sur une GC reste néanmoins un opérateur de généralisation : pour toute GC G et toute substitution σ, on a $L(G) \subseteq L(\sigma(G))$ et $FL(G) \subseteq FL(\sigma(G))$ (Buszkowski & Penn, 1990). Quand la cible de l'apprentissage est un langage régulier, l'usage est plutôt de le représenter par un automate fini. Or l'ensemble des automates finis obtenus à partir d'un automate fini quelconque A par l'opération de "fusion d'états" constitue, lui aussi, un treillis (Dupont et al., 1994). En définissant une correspondance entre GCs et automates finis, nous ouvrons la possibilité de "raccorder entre eux" des treillis jusqu'à présent distincts.

Les articles évoqués précédemment (Tellier, 2005c; Tellier, 2005b) montrent aussi que l'opération de "fusion d'état", qui est l'opérateur de généralisation usuel en inférence grammaticale régulière, peut être vu comme *un cas particulier de l'opération de substitution*. Plus précisément, les substitutions qui unifient les catégories affectées à un même mot de vocabulaire opèrent des "fusions de transitions" qui induisent elles-mêmes des fusions d'états. Mais, comme nous l'avons vu, cette opération de substitution peut faire sortir de l'espace des langages réguliers, alors que la fusion d'états, elle, transforme un automate fini en un autre automate fini. En quoi la "fusion de transitions" est-elle donc plus puissante ? L'Exemple 5 aide à le comprendre.

Exemple 5. *L'automate correspondant à la grammaire $FG(D)$ de l'Exemple 4 est donné en Figure 21.*

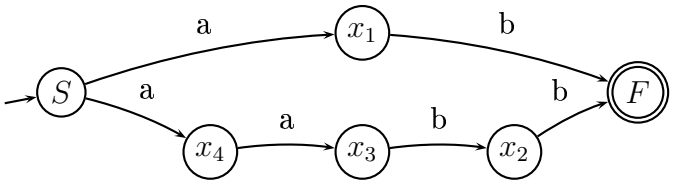

FIG. 5.8 – l'AF correspondant à $FG(D)$

La substitution définie par $\sigma(x_4) = \sigma(x_1) = x_3/x_2$, qui est unificatrice à la fois pour a et b, a pour effet :

- *d'unifier les états étiquetés x_1 et x_4 (par convention, l'état résultant est celui de plus petit indice, donc x_1) : c'est une fusion d'état classique ;*

– *de remplacer la transition entre x_3 et x_2 par une transition étiqueté par l'état x_1 : une transition étiquetée par un état fait référence au* langage de cet état, *c'est-à-dire à l'ensemble des chaînes correspondant à un chemin partant de cet état et aboutissant à un état final de l'automate. Cela peut donner lieu à des appels récursifs produisant des "peignes de peignes" comme on l'a vu Figure 20.*

L'automate généralisé représentant $\sigma(FG(D))$ est celui de la Figure 22.

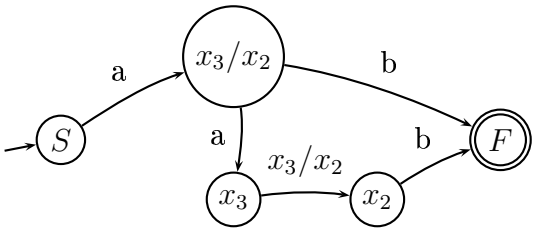

FIG. 5.9 – L'automate généralisé correspondant à $\sigma(FG(D))$

On peut rapprocher ce dispositif de celui des Réseaux de Transitions Récursifs ou RTRs (Woods, 1970), mais réduit à un seul automate (dans les RTRs, il y a autant d'automates que de symboles non terminaux). Celui de la Figure 22, comme on l'a déjà signalé, reconnaît $a^n b^n$.

Les liens entre GCs et ces cas particuliers de RTRs, que nous proposons d'appeler des "automates récursifs", ont été repris un peu plus en détail dans (Tellier, 2005a). L'espace des automates récursifs constitue un prolongement naturel de celui des automates finis, et permet d'envisager l'inférence de grammaires algébriques dans la continuité de l'inférence grammaticale régulière. De nombreux problèmes subsistent pourtant : on manque encore, pour les automates récursifs, d'un représentant canonique unique qui puisse servir de cible à un algorithme d'apprentissage. Ce rôle est joué, pour les langages réguliers, par l'automate déterministe minimal (Oncina & Garcia, 1992), ou encore l'AFER canonique (Denis et al., 2004).

Cette relation ouvre aussi des perspectives pour rapprocher des travaux réalisés indépendamment dans les deux espaces : nous pensons ainsi à l'apprentissage par données typées, évoqué dans le cadre des GCs en section 5.2, et envisagé dans celui des automates finis par Colin de la Higuera (Kermorvant & de la Higuera, 2002; Coste et al., 2004). Ces auteurs proposent un typage rudimentaire des états qui, en contraignant les fusions d'états possibles, améliore les techniques d'inférence grammaticale classiques. Les typages considérés dans les GCs s'appliquent, eux, aux mots du vocabulaire et donc aux transitions des automates correspondants. De plus, nous avons vu aussi que l'apprentissage par exemples typés est principalement une technique de *spécialisation*. Or, les les algorithmes classiques d'inférence grammaticale régulière implémentent presque tous des techniques de *généralisation*. La

seule tentative pour apprendre des automates finis par spécialisation en réalisant des "fissions d'états" est, à notre connaissance (Fredouille, 2000). Cette tentative n'était pas très concluante, car la spécialisation pure est combinatoirement trop explosive.

En adaptant notre technique d'apprentissage par exemples typés à des GCs représentant des automates finis, nous pensons être en mesure de proposer de nouvelles techniques d'apprentissage qui compléteraient celles déjà étudiées par le typage des états. Ce travail en cours n'a pas encore donné lieu à des publications, mais nous espérons dans les prochains mois être en mesure de rapprocher plus précisément ces deux perspectives.

5.3.3 Systèmes Question/Réponse et compositionnalité

Nous souhaitons achever ce parcours en évoquant un autre travail en cours qui, bien qu'apparemment éloigné des modèles présentés jusqu'à présent, entretient des liens avec lui. Il concerne des domaines où le Principe de compositionnalité n'est quasiment jamais explicitement invoqué, mais où il nous semble utile pour justifier certains travaux et en suggérer de nouveaux.

Les domaines sont l'extraction automatique à partir de textes, et les systèmes "Question/Réponse". En extraction automatique à partir de textes, il s'agit de sélectionner dans des corpus des mots ou des groupes de mots porteurs d'une certaine information, et susceptibles de remplir les champs d'un formulaire prédéfini (Poibeau, 2003). On peut penser, par exemple, à un corpus de dépêches d'agence de presse d'où il faut extraire les informations essentielles (les "qui", "quoi", "où", "quand", "comment" et "pourquoi" d'un article journalistique). Les systèmes Question/Réponse (désignés par la suite par "systèmes QR") sont encore plus ambitieux : ils se proposent de répondre à *n'importe quelle question factuelle*[4] formulée en langage naturel, en partant d'un corpus de textes ou en effectuant des requêtes sur le Web.

Pour réaliser ces tâches, la démarche est en partie commune ; elle consiste notamment à définir des "patterns d'extraction". On peut voir ces patterns comme des grammaires locales capables de sélectionner des portions de texte devant figurer juste avant et/ou juste après le(s) mot(s) intéressants à extraire. Ils peuvent prendre des formes très variées. L'apprentissage automatique est de plus en plus mis à contribution pour les apprendre automatiquement à partir de textes étiquetés. Dans notre exemple d'extraction, ces textes prendraient la forme de dépêches d'agence où les mots qui remplissent chaque champ du formulaire sont entourés de balises les identifiant.

[4] une question factuelle est une question qui requiert comme réponse une donnée correspondant à la valeur d'un champ d'un formulaire

Pour justifier théoriquement cette démarche, c'est l'*hypothèse de distributionnalité* qui est, le plus souvent, mise en avant. Cette hypothèse stipule que le sens d'un mot est en quelque sorte caractérisé dans un corpus par l'ensemble des contextes lexicaux dans lesquels il est employé. En identifiant ces contextes, on apprend à identifier des valeurs sémantiques. On suppose donc que le corpus disponible à l'apprentissage est suffisamment complet pour caractériser tous les contextes possibles.

Mais il nous semble que le Principe de compositionnalité est un moyen tout aussi pertinent de justifier ces travaux, surtout en présence de corpus incomplets ou au contraire trop grands pour être parcourus en entier (le Web), et qui obligent les techniques d'apprentissage à aller au-delà du simple recensement d'occurrences de mots et à opérer des *généralisations*. La vraie cible de ces systèmes, en effet (particulièrement des systèmes QR), c'est une *information sémantique*. La personne qui pose une question à un tel système ne cherche pas à savoir si le corpus disponible ou le Web contient une portion de texte d'une certaine forme : il attend une réponse. Certains systèmes QR cherchent d'ailleurs à effectuer des *inférences* pour répondre à des questions complexes.

Or les "patterns d'extraction" utilisés dans ces programmes portent, eux, exclusivement sur des informations lexicales et/ou syntaxiques : présence de certains mots, ou de mots d'une certaine catégorie grammaticale avant ou après le mot à extraire, régularités de constructions syntaxiques, etc. Utiliser des patterns syntaxiques pour chercher une information sémantique n'a de sens que si on fait l'hypothèse que la forme contraint le sens, autrement dit si on admet la pertinence du Principe de compositionnalité.

C'est dans cet esprit que nous avons abordé le domaine des systèmes QR, qui a constitué le sujet du stage de DEA de Florent Jousse en 2004, co-encadré avec Marc Tommasi (Jousse, 2004). Les systèmes QR "classiques" actuels sont en fait constitués d'un enchaînement de modules : un module de "typage" de la question et d'extraction de ses mots clés, un module de recherche d'information prenant appui sur ces mots clé pour sélectionner des textes ou des portions de textes susceptibles de contenir la réponse, et enfin un module d'extraction d'information tenant compte des résultats des deux premiers modules. Comme il était, bien sûr, impossible d'envisager l'écriture complète d'un tel système dans le cadre d'un DEA, nous nous sommes concentrés sur la phase ultime du système, celle qui se rapproche le plus d'une tâche d'extraction d'information à partir de textes. Nous avons, justement, cherché à clarifier les liens entre les deux tâches et nous avons montré que certaines techniques d'apprentissage automatique définies dans le cadre de l'extraction d'information (Califf, 1998; Ravichandran & Hovy, 2002; Marty & Torre, 2004) étaient réemployables avec profit dans le dernier module d'un système QR.

Ce travail a donné lieu à un article présenté à CORIA (Conférence française sur la Recherche d'Information et ses Applications) en 2005, et reproduit à la fin de ce document (Jousse et al., 2005). Il se poursuit en thèse actuellement.

Chapitre 6

Conclusion

Une grammaire formelle est un objet hautement structuré. Ce que nous avons essayé de montrer ici, c'est que son apprentissage n'est envisageable qu'à condition de disposer de données qui, elles aussi, sont *porteuses de structure*. Dans les travaux que nous avons exposés, nous avons en fait envisagé deux manières de faire passer un peu de la structure d'une analyse sntaxique complète dans des données d'apprentissage, par le biais d'un morphisme. Une analyse syntaxique complète, en effet, prend la forme d'un terme algébrique dont les noeuds sont des applications fonctionnelles et dont les feuilles sont des affectations de catégories à des mots du vocabulaire :
$FA(FA(\langle un, (S/(T\backslash S))/NC\rangle, \langle chat, NC\rangle), \langle dort, T\backslash S\rangle)$

Pour transformer ce terme en un *exemple structuré*, il suffit d'effacer les catégories figurant aux feuilles : $FA(FA(un,\ chat),\ dort)$. Pour transformer le même terme initial en un *exemple typé*, il suffit cette fois d'effacer les symboles d'applications fonctionnelles FA et BA, et d'appliquer le morphisme de transformation des catégories en types h aux feuilles : $\langle un, h((S/(T\backslash S))/NC)\rangle\langle chat, h(NC)\rangle\langle dort, h(T\backslash S))\rangle$.

Dans les deux cas, apprendre à partir de ces données revient donc à *remonter le sens d'un morphisme*. C'est ce que symbolisent aussi, à leur manière, les flèches de nos schémas des Figure 5.1 et 5.3. La validité des algorithmes d'apprentissage qui les utilisent en données d'entrée dépend, bien sûr, de conditions portant sur la grammaire initiale et/ou sur le morphisme, pour limiter les confusions et les ambiguités qu'il peut introduire :

- quand on efface les catégories, on ne peut apprendre que des GCs k-vlauées avec k connu à l'avance, c'est-à-dire des grammaires dont on peut prédire le nombre maximum de catégories différentes effacées ;
- quand on transforme les catégories en types, on ne peut apprendre que les grammaires dont on peut prévoir à l'avance combien de couples $\langle mot, type\rangle$ peuvent provenir d'une même affectation de catégorie à un mot (nous n'avons présenté ici que la condition plus restrictive où *un seul couple* $\langle mot, type\rangle$ est

permis par affectation, mais on peut l'étendre naturellelemnt à k couples par affectation, comme nous l'avons détaillé dans un article soumis).

De plus, nous avons montré que, dans les deux cas, la structure présente dans les données pouvait s'interpréter comme provenant de *connaissances sémantiques*. Toute notre argumentation se ramène donc au "slogan" suivant : *la sémantique, c'est de la structure (et inversement)*. Nous avons vu que cette "équation" pouvait s'interpréter au niveau des énoncés propositionnels comme au niveau du lexique. Evidemment, cette identification n'a été possible que grâce à une notion de "représentation des connaissances" à base de *formules bien formées*, qui elles-mêmes suivent les spécifications d'une *syntaxe formelle*, et à une version du Principe de compositionnalité adaptée au problème. Nous ne faisons donc ainsi que repousser à l'acquisition sémantique le problème de l'acquisition des structures portées par les formules logiques (ou par leur type).

Quelle que soit l'origine de ces structures, nous avons montré comment elles rendaient possible la conception d'algorithmes qui *spécialisent*. Sur ce point aussi, nous avons envisagé plusieurs alternatives : sélection *a posteriori* ou *a priori*. Plus la structure est prise en compte tôt, plus on a de chances de réduire la combinatoire. Jusqu'à présent, la plupart des programmes d'inférence grammaticale connus opéraient plutôt par *généralisation*. Les jeunes humains, pourtant, semblent plutôt fonctionner par spécialisation.

Nos hypothèses et nos postulats rejoignent largement ceux de la "linguistique cognitive", branche encore un peu marginale de la linguistique (Langacker, 1987; Lakoff, 1987; Heine, 1997), tout en continuant à se situer dans le cadre de l'informatique linguistique "classique". Mais nous sommes évidemment très consciente de l'écart qui subsiste entre les ambitions initiales annoncées dans le titre de ce document, qui visait la compréhension du mécanisme d'acquisition de leur langue maternelle par les enfants, et les résultats et algorithmes produits, qui sont des contributions essentiellement techniques et formelles. Les aspects expérimentaux n'ont pu être approfondis autant que cela aurait été souhaitable, en partie à cause de la difficulté de disposer de données adaptées. Les corpus étiquetés sémantiquement ne sont pas encore disponibles.

A défaut, les leçons tirées de ces recherches sont-elles applicables à d'autres domaines ? Le développement du Web et de formats de documents de type XML sont actuellement une source majeure de données structurées. Des projets de recherche commencent à émerger pour exploiter au mieux l'information portée par la structure de ces documents. C'est le cas du projet Mostrare de l'Inria, dont nous faisons partie, et qui a déjà donné lieu à des propositions intéressantes (Carme, 2005). Il est probable que les techniques et les modèles présentés ici (grammaires catégorielles, sémantique logique) ne sont pas directement transposables d'un domaine à un autre. Mais l'intuition sous-jacente au projet, à savoir que la structure

est porteuse d'information, est celle qui nous a menée jusque là.

Par ailleurs, la linguistique informatique s'est profondément renouvelée ces dernières années avec l'arrivée de techniques *fondées sur les données* plus que sur les modèles formels. On commence seulement maintenant à voir la possibilité de combiner les deux approches (Pereira, 2000). Les grammaires logiques lexicalisées sont probablement un des fomalismes le plus à même de s'"hybrider" avec des infomations statistiques (Villavicencio, 2002; Hockenmaier, 2005), et c'est une des pistes possibles que nous envisageons. Le travail exposé dans la toute dernière section de ce texte présente une autre façon de transposer les intuitions qui nous ont guidées jusque là dans un contexte plus expérimental. Enfin, même en s'en tenant aux modèles formels présentés ici, de nombreuses questions méritent approfondissement. Pour valider l'hypothèse de la primauté de la sémantique, il faudrait, en effet, être capable d'expliquer l'origine de tous les phénomènes syntaxiques (par exemple, ellipses, références pronominales) par un phénomène sémantique. C'est un chantier qui est à peine engagé.

Nous espérons avoir montré, tout au long de ce texte, la grande cohérence des thèmes des recherche que nous avons menées ces dernières années, ainsi que celle des moyens employés pour les aborder. Le "champ parcouru" au final n'est peut-être pas très vaste, les sillons creusés sont souvent proches et parallèles et la récolte est modeste[1]. Au moins y avons-nous trouvé beaucoup de plaisir.

[1]précisons d'ailleurs que, pendant cette période, nous avons aussi contribué à l'encadrement de deux autres stagiaires de DEA (dont l'un qui continue en thèse depuis 2 ans) sur des sujets très différents que nous avons préféré ne pas aborder dans ce document pour préserver sa thématique

Bibliography

Adriaans, P. W. (1992). *Language Learning from a Categorial Perspective.* Doctoral dissertation, University of Amsterdam, Amsterdam, The Netherlands.

Anderson, J. R. (1977). Induction of augmented transition networks. *Cognitive Science*, *1*, 125–157.

Angluin, D. (1980a). Finding patterns common to a set of strings. *Journal of Computer and System Sciences*, *21*, 46–62.

Angluin, D. (1980b). Inductive inference of formal languages from positive data. *Inform. Control*, *45*, 117–135.

Angluin, D. (1982). Inference of reversible languages. *J. ACM*, *29*, 741–765.

Angluin, D. (1987). Learning regular sets from queries and counterexamples. *Information and Computation*, *75*, 87–106.

Angluin, D. (1988). Queries and concept learning. *Machine Learning*, *2*, 319–342.

Angluin, D. (1992). Computational learning theory : Survey and selected bibliography. *Proceedings of the 24th Annual ACM Symposium on the Theory of Computing* (pp. 351–369).

Arsac, J. (1987). *Les machines à penser : des ordinateurs et des hommes.* Le Seuil.

Arsac, J., & Vauthier, J. (1989). *Jacques arsac, un informaticien : entretien avec jacques vauthier.*

Asher, N. (1993). *Reference to abstract objects in discourse.* Kluwer.

Bar Hillel, Y., Gaifman, C., & Shamir, E. (1960). On categorial and phrase structure grammars. *Bulletin of the Research Council of Israel*, *9F*.

Bargelli, D., & Lambek, J. (2001). An algebraic approach to french sentence structure. *proceedings of the 4th conference LACL* (pp. 62–78). Springer Verlag.

Barwise, J., & Perry, J. (1983). *Situation and attitudes.* MIT Press.

Besombes, J., & Marion, J.-Y. (2004). Learning reversible categorial grammars from structures. *Categorial Gramars.*

Beust, P. (1998). *Contribution à un modèle intéractionniste du sens.* Doctoral dissertation, université de Caen.

Bickerton, D. (1990). *Language and species.* "The University of Chicago Press".

Blache, P. (2001). *Les grammaires de propriétés.* Hermès.

Bonato, R., & Rétoré, C. (2001). Learning rigid Lambek grammars and minimalist grammars from structured sentences. *Proceedings of LLL01* (pp. 23–34).

Brent, M. R. (1996). *Computational approaches to language acquisition.* MIT Press.

Buszkowski, W. (2001). Lambek grammars based on pregroups. *proceedings of LACL01* (pp. 95–109). Springer Verlag.

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

Béchet, D., Dikovsky, A., & Foret, A. (2005). Dependancy structure grammars. *proceedings of LACL 05* (pp. 18–34). Springer Verlag.

Béchet, D., & Foret, A. (2003). Remarques et perspectives sur les langages de prégroupes d'ordre 1/2. In *proceedings of taln 03 (poster).*

Béchet, D., Foret, A., & Tellier, I. (2004). Learnability of pregroup grammars. *7th International Colloquium on Grammatical Inference* (pp. 65–76). Springer Verlag.

Califf, M. (1998). *Relational learning techniques for natural language information extraction* (Technical Report AI98-276). IA Laboratory, university of Texas of Austin.

Carme, J. (2005). *Inférence de requêtes dans les arbres et applications à l'extraction d'informations sur le web.* Doctoral dissertation, Université Charles-de-Gaulle - Lille 3.

Chambreuil, M. (1989). *Grammaire de montague ; langage, traduction, interprétation.* Adossa.

Chomsky, N. (1957). *Syntactic structures.* The Hague.

Chomsky, N. (1968). *Language and mind.* Brace and World.

Chomsky, N. (2001). *Règles et représentations.* Flammarion.

Christophe, A. (2002). L'apprentissage du langage : une capacité innée ? *Intellectica*, 189–210.

Claveau, V., & Sebillot, P. (2004). Extension de requêtes par lien sémantique nom-verbe acquis sur corpus. *proceedings of TALN 2004* (pp. 121–130).

Claveau, V., & Sébillot, P. (2004). Apprentissage semi-supervisé de patrons d'extraction de couples nom-verbe. *TAL, 45.*

Cornuéjols, A., & Miclet, L. (2002). *Apprentissage artificiel ; concepts et algorithmes.* Eyrolles.

Coste, F., Fredouille, D., Kermovant, C., & de la Higuera, C. (2004). Introducing domain and typing bias in automata inference. *proceedings of the 7th ICGI* (pp. 115–126). Springer Verlag.

Crevier, D. (1999). *A la recherche de l'intelligence artificielle.* Flammarion.

Cyrulnik, B. (1995). *La naissance du sens.* Hachette.

Daelemans, W. (2005). *Memory-based language processing.* Cambridge University Press.

de Boysson-Bardie, B. (1999). *Comment la parole vient aux enfants.* Odile Jacob.

de Groote, P. (2001). Towards abstract categorial grammars. *proceedings of ACL 2001* (pp. 148–155).

de Groote, P., & Pogodalla, S. (2004). On the expressive power of abstract categorial grammars : Representing context-free formalisms. *Journal of Logic, Language and Information, 4*, 421–438.

de la Higuera, C. (1997). Characteristic sets for polynomial grammatical inference. *Machine Learning, 27*, 125–137.

de Saussure, F. (1916). *Cours de linguistique générale.* Payot.

Delsarte, P., & Thayse, A. (2001). *Logique pour le traitement de la langue naturelle.* Hermès.

Denis, F., D'Halluin, C., & Gilleron, R. (1996). PAC learning with simple examples. *13th Annual Symposium on Theoretical Aspects of Computer Science* (pp. 231–242). Grenoble, France : Springer Verlag.

Denis, F., & Gilleron, R. (2001). Pac learning under helpful distributions. *Theoretical Informatics and Applications, 35*, 129–148.

Denis, F., Lemay, A., & Terlutte, A. (2004). Learning regular languages using rfsas. *Theoretical Computer Science, 313*, 267–294.

Desclés, J.-P. (1982). *Penser les mathématiques*, chapter Quelques réflexions sur les rapports entre linguistique et mathématiques, 88–107. Le Seuil.

Dessalles, J.-L. (2000). *Aux origines du langage.* Hermès.

Dikovsky, A. (2000). Dependencies on the other side of the curtain. *TAL*, 79–111.

Dikovsky, A. (2003). Linguistic meaning from the language acquisition perspective. *proceeding of the 8th conference on Formal Grammars* (pp. 63–76).

Dikovsky, A. (2004). From prosodic patterns to basic semantic types : A bootstrapping hypothesis. *proceedings of the 7th International Conference on Cognitive Modeling in Linguistics (CML2004)* (pp. 174–182).

Dikovsky, A. (2005). Underspecified semantics for dependency grammars. *Proceedings of the 4th Mexican International Conference on Artificial Intellience* (pp. 741–751). Springer Verlag.

Dowty, D. R., Wall, R. E., & Peters, S. (1981). *Introduction to montague semantics.* Linguistics and Philosophy. Reidel.

Dudau, D. (2000). Apprentissage du langage naturel. Master's thesis, Université Lille1.

Dudau-Sofronie, D., & Tellier, I. (2004a). A study of learnability of lambek grammars from typed examples. *Proceedings of Categorial Grammars 04* (pp. 133–147). Montpellier, France.

Dudau-Sofronie, D., & Tellier, I. (2004b). Un modèle d'acquisition de la syntaxe à l'aide d'informations sémantiques. *actes de la 11ème Conférence TALN, Traitement Automatique du Langage Naturel* (pp. 137–146). Fès, Maroc.

Dudau-Sofronie, D., Tellier, I., & Tommasi, M. (2001a). From logic to grammars via types. *proceedings of LLL 2001, Learning Language in Logic* (pp. 35–46).

Dudau-Sofronie, D., Tellier, I., & Tommasi, M. (2001b). Learning categorial grammars from semantic types. *proceedings of the 13th Amsterdam Colloquium* (pp. 79–84).

Dudau-Sofronie, D., Tellier, I., & Tommasi, M. (2002). A tool for language learning based on categorial grammars and semantic information. *proceedings of ICGI'2002, International Colloquium on Grammatical Inference (demo session)* (pp. 303–305). Springer Verlag.

Dudau-Sofronie, D., Tellier, I., & Tommasi, M. (2003a). A learnable class of classical categorial grammars from typed examples. *8th Conference on Formal Grammar* (pp. 77–88).

Dudau-Sofronie, D., Tellier, I., & Tommasi, M. (2003b). Une classe de grammaires catégorielles apprenable à partir d'exemples typés. *5ème Conférence francophone sur l'apprentissage automatique* (pp. 169–184). Presses Universitaires de Grenoble.

Dupont, F. (2003). Apprentissage de grammaires de lambek à partir d'exemples typés. Master's thesis, Université Lille1.

Dupont, P., Miclet, L., & Vidal, E. (1994). What is the search space of the regular inference. *ICGI'94 - Lectures Notes in Computer Science* (pp. 25–37). Heidelberg.

Dupuy, J. P. (1994). *Aux origines des sciences cognitives.* La Découverte.

Emms, M. (1994). Extraction covering extensions of lambek calculus are not cf. *proceedings of the 9th Amsterdam Colloquium.*

Enard, W., Przeworski, M., Fisher, S., Lai, C., Wiebe, V., Kitano, T., Monaco, A., & Paabo, S. (2002). Molecular evolution of foxp2, a gene involved in speech and language. *Nature*, 869–872.

Erk, K., Koller, A., & Niehren, J. (2002). Processing underspecified semantic representations in the constraint language for lambda structures. *Journal of Research on Language and Computation*, *1*, 127–169.

Fauconnier, G. (1984). *Espaces mentaux; aspects de la construction du sens dans les langues naturelles.* Minuit.

Feldman, J. A. (1998). Real language learning. *ICGI'98, 4th International Colloquium in Grammatical Inference* (pp. 114–125).

Florêncio, C. C. (2002). Consistent identification in the limit of rigid grammars from strings is np-hard. *Grammatical Inference : Algorithms and Applications* (pp. 49–62). Springer Verlag.

Florin, A. (1999). *Le développement du langage.* Dunod.

Florêncio, C. C. (2000). On the complexity of consistent identification of some classes of structure languages. *ICGI'2000, 5th International Colloquium on Grammatical Inference* (pp. 89–102). Springer Verlag.

Florêncio, C. C. (2001). Consistent identification in the limit of any of the classes k-valued is NP-hard. *Logical Aspects of Computational Linguistics* (pp. 125–134). Springer Verlag.

Florêncio, C. C. (2003). *Learning categorial grammars.* Doctoral dissertation, Utrecht University.

Fodor, J. (1975). *The language of thought.* Harvester Press.

Foret, A., & le Nir, Y. (2002). On limit points for some variants of rigid lambek grammars. *Grammatical Inference : Algorithms and Applications* (pp. 106–119). Springer Verlag.

Fredouille, D. (2000). Expériences sur l'inférence de langage par spécialisation. *proceedings of CAP'2000* (pp. 117–130).

Fulop, S. (2005). Semantic bootstrapping of type-logical grammar. *Journal of Logic, Language and Information*, *14*, 49–86.

Galmiche, M. (1991). *sémantique linguistique et logique.* Presses Universitaires de France.

Gardner, H. (1993). *Histoire de la révolution cognitive.* Payot.

Gernsbacher, M. (1990). *Language comprehension as structure building.* Lawrence Erlbaum.

Gest, J. (1998). Apprentissage syntaxico-sémantique du langage naturel. mémoire de dea, Université Lille 1.

Gold, E. (1967). Language identification in the limit. *Inform. Control*, *10*, 447–474.

Goldman, S. A., & Mathias, H. D. (1996). Teaching a smarter learner. *Journal of Computer and System Sciences*, *52*, 255–267.

Grimshaw, J. (1981). *The logical problem of language acquisition.* MIT Press.

Groenendijk, J., & Stockhof, M. (1991). *Dynamic predicate logic*, vol. 3 of *Linguistics and Philosophy*, 175–204. Kluwer.

Groenendijk, J., Stockhof, M., & Veltman, F. (1996a). Changez le contexte ! *Languages.*

Groenendijk, J., Stockhof, M., & Veltman, F. (1996b). *Handbook of contemporary semantic theory*, chapter Coreference and modality, 179–213.

Gruselle, J.-P. (1997). *Le rôle du mot dans la formation des concepts : modèle informatique et son implémentation.* Doctoral dissertation, université Paris 11-LIMSI.

Hamburger, H., & Wexler, K. (1975). A mathematical theory of learning transformational grammar. *Journal of Mathematical Psychology, 12*, 137–177.

Heine, B. (1997). *Cognitive foundation of grammar.* Oxford University Press.

Hirsh-Pasek, K., Tucker, M., & Golinkoff, R. M. (1995). *Signal to syntax : Bootstrapping from speech to grammar in early acquisition.* Lawrence Erlbaum.

Hobbs, J., & Rosenstein, S. J. (1977). Making computational sense of montague's intensional logic. *Artificial Intelligence*, 287–306.

Hockenmaier, J. (2005). *Data and models for statistical parsing with combinatory categorial grammar.* Doctoral dissertation, University of Edinburg.

Hodges, A. (1983). *Alan turing ou l'énigme de l'intelligence.* Payot.

Houdé, O. (1998). *Rationnalité, développement et inhibition.* Presses Universitaires de France.

Israel, D. (2003). The very idea of dynamic semantics. *proceedings of the 9th Amsterdam Colloquium.*

Jackendoff, R. (1990). *Semantic structures.* MIT Press.

Janssen, T. M. V. (1997). Compositionality. In J. V. Benthem and ter A. Meulen (Eds.), *Handbook of logic and language*, 417–473. MIT Press.

Johnson-Laird, P. N. (1983). *Mental models.* Cambridge University Press.

Joshi, A., & Schabes, Y. (1997). *Handbook of formal languages, vol3*, chapter Tree-Adjoining Grammars, 69–120. Springer Verlag.

Joshi, A. K., Kallmeyer, L., & Romero, M. In *proceedings of the 5th international workshop on computational semantics (iwcs-5), title =.*

Jousse, F. (2004). Apprentissage automatique de motifs d'extraction pour les systèmes question/réponse. Master's thesis, université Lille1.

Jousse, F., Tellier, I., Tommasi, M., & Marty, P. (2005). Learning to extract answers in question answering : Experimental studies. *Actes de CORIA'05* (pp. p.85–100). Hermès.

Kamp, H., & Reyle, U. (1993). *From discourse to logic; introduction to the modeltheoretic semantics of natural language.* Reidel.

Kanazawa, M. (1996). Identification in the limit of categorial grammars. *Journal of Logic, Language and Information, 5*, 115–155.

Kanazawa, M. (1998). *Learnable classes of categorial grammars.* The European Association for Logic, Language and Information. CLSI Publications.

Kanazawa, M. (2001). Learning word-to-meaning mappings in logical semantics. *Amsterdam Colloquium 2001* (pp. 126–131).

Kandel, E. R., Schwartz, J. H., & Jessel, T. M. (1995). *Essentials of neural science and behavior.* Prentice-Hall.

Kaplan, F. (2001). *La naissance d'une langue chez les robots.* Hermès.

Kayser, D. (1997). *La représentation des connaissances.* Hermès.

Kermorvant, C., & de la Higuera, C. (2002). Learning language with help. *6th International Colloquium on Grammatical Inference* (pp. 161–173). Springer Verlag.

Kirby, S. (2002). *Learning, bottlenecks and the evolution of recursive syntax*, chapter 6, 173–204. Cambridge University Press.

Koller, A., Niehren, J., & Striegnitz, K. (2000). Relaxing underspecified semantic representations for reinterpretation. *Grammars, 3*, 217–241. Special Issue on MOL'99.

Lai, C., Fisher, S., Hurst, J., Vargha-Khadem, F., & Monaco, A. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature, 4*, 465–456.

Lakoff, G. (1987). *Women, fire and dangerous things.*

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

Lambek, J. (1997). Type grammars revisited. *proceedings of LACL 97.* Springer Verlag.

Lambek, J. (2001). Type grammars as pregroups. *Grammars, 4*, 21–39.

Lambek, J., & Casadio, C. (2001). An algebraic analysis of clitic pronouns in italian. *proceedings of the 4th Conference LACL* (pp. 110–124). Springer Verlag.

Lambek, J., & Casadio, C. (2002). A tale of four grammars. *Studia Logica, 71*, 315–329.

Langacker, R. W. (1987). *Foundation of cognitive grammar.* Stanford University Press.

Langlet, P. (1982). Language acquisition through error discovery. *Cognition and Brain Theory*, 211–255.

Manning, C. (1999). *Foundations of statistical natural language processing.* MIT Press.

Martinet, A. (1960). *Elements de linguistique générale.* Armand Colin, Paris.

Marty, P., & Torre, F. (2004). Codages et connaissances en extraction d'information. *6ième Conférence francophone sur l'apprentissage automatique* (pp. 207–222). Presses Universitaires de Grenoble.

Mehler, J., & Dupoux, E. (1995). *Naître humain.* Odile Jacob.

Messiant, C. (2005). La modélisation de l'acquisition de la sémantique lexicale vue à travers les travaux de j. m. siskind. Master's thesis, DEA sciences cognitives, université Lille3.

Michaelis, J. (2001). Transforming linear context-free rewriting systems into minimalist grammars. *proceedings of LACL 01* (pp. 228–244). Springer Verlag.

Michaelis, J., & Kracht, M. (1996). Semilinearity as a syntactic invariant. *proceedings of LACL 96* (pp. 329–345).

Michalski, R., Carbonell, J., & Mitchell, T. (Eds.). (1986). *Machine learning : An artificial intelligence approach*, vol. II. Los Altos, California : Morgan Kaufmann.

Miller, P. (1999). *Strong generative capacity.* CLSI Publications.

Miller, P., & Torris, T. (1990). *Formalismes syntaxiques pour le traitement automatique du langage naturel.* Hermès.

Mitchell, T. (1997). *Machine learning.* McGraw-Hill.

Montague, R. (1974). *Formal philosophy; selected papers of richard montague.* Yale University Press.

Moortgat, M. (1988). *Categorial investigations : Logical and linguistics aspects of the lambek calculus.*

Moortgat, M. (1997). *Handbook of logic and language*, chapter Categorial type logics. Elsevier.

Moot, R. (2003). Parsing corpus-induced type-logical grammars. *proceedings of of the CoLogNet/ElsNet Workshop on Linguistic Corpora and Logic Based Grammar Formalisms.*

Moreau, E. (2004). Apprentissage partiel de grammaires catégorielles. *TALN 2004* (pp. 299–308).

Morill, G. (1994). *Type logical grammar.* Kluwer.

Muskens, R. (1994). A compositional discourse representation theory. *proceedings of the 9th Amsterdam colloquium* (pp. 467–486).

Nazarenko, A. (1998). *Tal, numéro spécial sur la compositionnalité*, vol. 39. Hermès.

Nyckees, V. (1998). *La sémantique.* Belin.

Oehrle, R. T., Bach, E., & Wheeler, D. (Eds.). (1988). *Categorial grammars and natural language structures.* Dordrecht : D. Reidel Publishing Company.

Oncina, J., & Garcia, P. (1992). Inferring regular languages in polynomial update time. *Pattern Recognition and Image Analysis* (pp. 49–61).

Osborne, M., & Briscoe, T. (1998). Learning stochastic categorial grammars. *CoNLL97 : Computational Natural Language Learning* (pp. 80–87).

Partee, B. (1990). *Mathematical methods in linguistics.* No. 30 in Linguistics and Philosophy. Kluwer.

Pentus, M. (1993). Lambek grammars are context-free. *proceedings of the 8th annual IEE Symposium on Logic in Computer Science* (pp. 429–433).

Pereira, F. (2000). Formal grammar and information theory : Together again ? *Philosophical Transactions of the Royal Society, 358 (1769)*, 1239–1253.

Piaget, J. (2003). *La représentation du monde chez l'enfant.* Presses Universitaires de France.

Piatelli-Palmarini, M. (1979). *Théories du langage, théories de l'apprentissage, le débat entre jean piaget et noam chomski.* Le Seuil.

Pinker, S. (1984). *Language learnability and language development.* Harvard University Press.

Pinker, S. (1994). *The language instinct.* Penguin Press.

Pinker, S. (1995). *An invitation to cognitive science*, chapter Language Acquisition, 135–182. MIT Press.

Pinker, S. (2000). *Comment fonctionne l'esprit.* Odile Jacob.

Pitt, L. (1989). Inductive Inference, DFAs, and Computational Complexity. *Proceedings of AII-89 Workshop on Analogical andInductive Inference ; Lecture Notes in Artificial Intelligence 397* (pp. 18–44). Heidelberg : Springer-Verlag.

Pogodalla, S. (2001). *Réseaux de preuve et génération pour les grammaires de types logiques.* Doctoral dissertation, Institut National Polytechnique de Lorraine.

Pogodalla, S. (2004). Using and extending ACG technology : Endowing categorial grammars with an underspecified semantic representation. *proceedings of Formal Grammars* (pp. 197–209).

Poibeau, T. (2003). *Extraction automatique d'information.* Hermès.

Popescu-Belis, A. (1999). *Modélisation multi-agent des échanges langagiers : application au problème de la référence et à son évaluation.* Doctoral dissertation, Paris XI, LIMSI.

Popper, K. (1989). *La quête inachevée.* Press Pocket.

Preller, A. (2005). Category theoretical semantics for pregroup grammars. *proceedings of LACL 05* (pp. 236–254). Springer Verlag.

Pustejovsky, J. (1995). *The generative lexicon.* MIT Press.

Quillian, R. (1968). *Semantic information processing*, chapter Semantic Memory. MIT Press.

Radzinski, D. (1991). Chinese number-names, tree adjoining grammars and mild context-sensitivity. *Computational Linguistics, 17*, 277–299.

Rastier, F., Cavazza, M., & Abeillé, A. (1997). *Sémantique pour l'analyse.* Masson.

Ravichandran, D., & Hovy, E. (2002). Learning surface text patterns for a question answering system. *Proceedings of the ACL conference.*

Reyle, U. (1995a). On reasoning with ambiguities. *proceedings of the 6th Meeting of the Association for Computational Linguistics* (pp. 1–8).

Reyle, U. (1995b). Underspecified discourse representation structures and their logic. *Bulletin of the Interest Group on Logic Programming.*

Rétoré, C. (2002). Logique linéaire et syntaxe des langues. Habilitation à diriger des recherches.

Sabah, G. (1990). *L'intelligence artificielle et le langage*, vol. tome 1 : représentation des connaissances. Hermès.

Sakakibara, Y. (1990). Learning context-free grammars from structural data in polynomial time. *Theoretical Computer Science, 76*, 223 – 242.

Sakakibara, Y. (1992). Efficient learning of context-free grammars from positive structural examples. *Information and Computation, 97*, 23–60.

Schank, R., & Abelson, R. (1977). *Scripts, plans, goals and understanding.* Lawrence Erlbaum.

Schieber, S. (1985). *Evidence against the context-freeness of natural language*, vol. 8 of *Linguistics and Philosophy*, 333–343. Springer Verlag.

Searle, J. (1980). Minds, brains, and programs. *Bulletin of the European Association of Theoretical Computer Science*, 417–458.

Siskind, J. M. (1996). *Computational approaches to language acquisition*, chapter A computational study of cross-situational techniques for learning word-to-word mappings. MIT Press.

Sofronie, D. D. (2004). *Apprentissage de grammaires catégorielles pour simuler l'acquisition du langage naturel à l'aide d'informations sémantiques.* Doctoral dissertation, Université Lille 1.

Sowa, J. (1984). *Conceptual structures.* Addison-Wesley.

Stabler, E. (2001). *Linguistic form and its computation*, chapter Minimalist grammars and recognition, 327–352. CLSI Publications.

Steedman, M. (1996). *Surface structure and interpretation.* MIT Press.

Steedman, M. (2000). *the syntactic process.* MIT Press.

Steedman, M., & alii (2002). *Semi-supervised training for statistical parsing* (Technical Report). CLSP WS.

Tellier, I. (1992). Système de lambek étendu pour la traduction logique de phrases en langage naturel. Master's thesis, ENSEEIHT-informatique.

Tellier, I. (1998a). Apprentissage syntaxico-sémantique du langage naturel. *actes des 13èmes JFA'98, Journées Francophones sur l'Apprentissage* (pp. 13–25).

Tellier, I. (1998b). Meaning helps learning syntax. *proceedings of ICGI'98, International Colloquium on Grammatical Inference* (pp. 25–36). Springer Verlag.

Tellier, I. (1998c). Syntactico-semantic learning of categorical grammars. *proceedings of NeMLaP3/CoNLL98 ACL Workshop on Paradigms and Grounding in Language Learning* (pp. 311–314).

Tellier, I. (1999a). *Learning to understand* (Technical Report IT-320). Laboratoire d'Informatique Fondamentale de Lille.

Tellier, I. (1999b). Rôle de la compositionnalité dans l'acquisition d'une langue. *actes de CAP'99, 1ère Conférence d'APprentissage* (pp. 107–114).

Tellier, I. (1999c). Towards a semantic-based theory of language learning. *proceedings of the 12th Amsterdam Colloquium* (pp. 217–222).

Tellier, I. (2000). Semantic-driven emergence of syntax : the principle of compositionality upside-down. *proceedings of Evolang 2000, 3rd conference on the Evolution of Language* (pp. 220–224).

Tellier, I. (2005a). Automata and ab-categorial grammars. *CIAA 05 (1Oth International Conference on Implementation and Application of Automata)* (pp. 287–288).

Tellier, I. (2005b). Inférence grammaticale et grammaires catégorielles : vers la grande unification ! *7ème Conférence en Apprentissage* (pp. 63–78). Presses Universitaires de Grenoble.

Tellier, I. (2005c). When categorial grammars meet regular grammatical inference. *5th International Conference on Logical Aspects of Computational Linguistics* (pp. p.317–332). Springer Verlag.

Thompson, C. A., Mooney, R. J., & Tang, L. R. (1997). Learning to parse natural language database queries. *Machine Learning Workshop on Automata Induction, Grammatical Inference and Language Acquisition.*

Turing, A. (1950). *Computing machinery and intelligence.*

Valiant, L. (1984). A theory of the learnable. *Commun. ACM, 27*, 1134–1142.

Villavicencio, A. (2002). *The acquisition of a unification-based generalised categorial grammar.* Doctoral dissertation, university of Cambridge.

Wehrli, E. (1997). *L'analyse syntaxique des langues naturelles.* Masson.

Wexler, K., & Culicover, P. (1980). *Formal principles of language acquisition.* MIT Press.

Winston, P. (1995). Learning structural descriptions from examples. *Psychology of Computer Vision.*

Wolper, P. (1997). *Introduction à la calculabilité.* InterEditions.

Woods, W. A. (1970). Transition network grammars of natural language analysis. *Communications of the ACM*, 591–606.

Yokomori (1995). On polynomial-time learnability in the limit of strictly deterministic automata. *Machine Learning, 2*, 153–179.

Chapitre 7

Articles sélectionnés

Les articles reproduits dans le suite de ce document sont les suivants :

- I. Tellier : "Meaning Helps Learning Syntax" actes de ICGI'98 (4th International Colloquium on Grammatical Inference), Lecture Notes in Artificial Intelligence 1433, p.25-36, 1998.
- I. Tellier : "Towards a Semantic-based Theory of Language Learning", actes du 12th Amsterdam Colloquium, p217-222, 1999.
- D. Dudau-Sofronie, I. Tellier et M. Tommasi : "Une classe de grammaires catégorielles apprenable à partir d'exemples typés", actes de CAP'03, 5ème Conférence francophone sur l'apprentissage automatique, p.169-184, 2003.
- D. Béchet, A. Foret et I. Tellier : "Learnability of Pregroup Grammars", actes de ICGI'04 (7th International Colloquium on Grammatical Inference), Lecture Notes in Artificial Intelligence 3264, p.65-76, 2004.
- I. Tellier : "When Categorial Grammars meet Regular Grammatical Inference", 5th International Conference LACL (Logical Aspects of Computational Linguistics), Lecture Notes in Artificial Intelligence 4492, p.317-332, 2005.
- F. Jousse, I. Tellier, M. Tommasi et P. Marty : " Learning to Extract Answers in Question Answering : Experimental Studies", Actes de CORIA'05, p.85-100, 2005.