Modèles algorithmiques de l'acquisition de la syntaxe : concepts et méthodes, résultats et problèmes

Denis Bechet, Roberto Bonato, Alexandre Dikovsky, Annie Foret, Yannick Le Nir, Erwan Moreau, Christian Retoré, Isabelle Tellier

Denis Bechet[1], Roberto Bonato[2], Alexandre Dikovsky[1], Annie Foret[3], Yannick Le Nir[4], Erwan Moreau[1], Christian Retoré[5], Isabelle Tellier[6]
1 : LINA, Université de Nantes : prenom.nom@univ-nantes.fr
2 : DeepBlue, Roma : robonato@gmail.com,
3 : IRISA, Université de Rennes 1 : foret@irisa.fr
4 : INRIA-Futurs, LIUPPA & EISTI, Pau : yannick.lenir@eisti.fr
5 : INRIA-Futurs & LaBRI, Université Bordeaux 1 retore@labri.fr
6 : INRIA-Futurs, Unviersité de Lille 3 : isabelle.tellier@univ-lille3.fr

MODÈLES ALGORITHMIQUES DE L'ACQUISTION DE LA SYNTAXE : concepts et méthodes, résultats et problèmes

Résumé

Dans cet article, nous présentons nos résultats récents concernant l'apprentissage de la syntaxe des langues naturelles, en adoptant le point de vue de l'inférence grammaticale symbolique. L'objectif est d'identifier à partir d'exemples, dans une classe de grammaires connue à l'avance, une grammaire particulière qui engendre les dits-exemples. Le modèle de Gold fixe les conditions et le critère de réussite d'une telle entreprise : quand un algorithme produisant une grammaire candidate existe-t-il ? quelle structure doivent contenir les exemples : suites de mots, suites de mots étiquetés, arbres d'analyse ? D'un point de vue théorique, nos résultats établissent l'apprenabilité ou la non-apprenabilité de certaines classes de grammaires catégorielles. En pratique, nos résultats permettent aussi d'acquérir automatiquement des ressources syntaxiques à partir de données réelles. Au final, nous discutons de l'intérêt de cette approche pour modéliser l'acquisition de leur langue naturelle par les enfants ainsi que pour construire automatiquement des grammaires électroniques à partir de corpus.
Mots-clés: acquisition syntaxique, inférence grammaticale, grammaires catégorielles, modèle de Gold, ressources syntaxiques.

Abstract

In this paper, we present our recent results on the acquistion of the syntax of natural languages, from the point of view of the theory of grammatical inference. Given a class of possible grammars, the objective is to identify, from a set of positive examples, a grammar in the class which produces the examples. The Gold model formalises the learning process and gives stringent criteria of its success: when does there exist an algorithm producing a target grammar ? what kind of structure should the examples have (strings of words, strings of tagged words, trees) ? From a theoretical point of view, our results establish the learnability or the unlearnability of various classes of categorial grammars. From a practical perspective, these results enable the extraction of syntactic information from real data. Finally, we discuss the interest of this approach for modelling child language acquisition and for automated induction of grammars from corpora.
Key-words: syntax learning, grammatical inference, categorial grammars, Gold's model, syntactical resources.

2 1. Présentation

La plupart des travaux utilisant une approche inductive à des fins linguistiques extraient et exploitent les régularités statistiques des corpus. Les structures sont rarement prises en compte et, lorsqu'elles le sont, ce sont toujours des structures incomplètes et en partie erronées.

Notre article relève d'une autre approche qu'on pourrait qualifier d'exacte : c'est celle de l'apprentissage automatique symbolique, plus précisément de l'« inférence grammaticale ». Dans une approche statistique, l'acquisition de la grammaire est finie lorsque, pour suffisamment de phrases, la ou les bonnes analyses figurent parmi celles proposées par la grammaire acquise. En inférence grammaticale, le critère de succès est objectif et précis : la grammaire construite doit engendrer exactement le langage énuméré si suffisamment d'exemples lui ont été présentés. Les questions que nous nous posons sont donc les suivantes : dans quelles conditions existe-t-il un algorithme qui puisse, à partir d'exemples de phrases syntaxiquement correctes, identifier une grammaire formelle qui les produise ? Lorsqu'il existe, comment fonctionne-t-il, quelles sont ses propriétés ? Ce domaine a deux motivations principales : la modélisation de l'apprentissage de leur langue maternelle par les enfants et l'acquisition automatique ou semi-automatique de grammaires électroniques. Commençons donc par détailler les motivations de cette approche exacte.

1.1. Acquisition naturelle

Lors de ses trois premières années de vie, tout enfant apprend, sans aucun effort particulier, à comprendre et à parler sa langue maternelle. C'est un véritable miracle car les langues sont des systèmes symboliques d'une grande complexité. S'il est probable que certains principes universels soient innés (Pinker 1994), la diversité des langues empêche que la totalité du système linguistique soit génétiquement codée.

L'apprentissage de la langue maternelle nécessite un contact linguistique direct avec les parents ainsi qu'une interaction avec l'environnement : par exemple, la télévision ne suffit pas. Etonnamment, lors des quatre premières années, les corrections des tuteurs n'ont pas d'influence : avec ou sans elles, l'acquisition suit les mêmes étapes, dans les mêmes délais et avec la même réussite (Pinker 1994).

On distingue désormais classiquement trois phases dans ce processus (Hirsh-Pasek, Golinkoff 1999, Boisson-Bardie 1996). Durant la phase I (0-9 mois) se déroulent le « packaging » et la segmentation acoustique : l'enfant établit un lien entre les événements de son entourage et la perception de

certains signaux acoustiques dans les paroles qu'on lui adresse. La phase II (9-24 mois) est celle de l'acquisition lexicale : l'enfant apprend ses premiers mots et forme des structures argumentales simples. À l'issue de cette phase, ses phrases comportent deux ou trois mots (un verbe, un sujet ou/et un objet) mais il ne sait pas encore les composer. C'est durant la phase III (24-36 mois), qu'il acquiert enfin syntaxe et morphologie. Même si ces dernières continuent à légèrement évoluer par la suite, on peut alors considérer qu'il maîtrise sa langue. Les principaux facteurs qui guident l'enfant varient d'une phase à l'autre (en phase I la prosodie, en phase II la sémantique et en phase III la syntaxe) mais, dans tous les cas, ce sont les expériences de communication langagière réussies qui sont essentielles. Nous verrons plus loin quelle incidence cette observation aura sur les modèles d'apprentissage que nous considérerons.

Après un demi-siècle de recherches, en dépit de l'énorme quantité de faits observés et systématisés dans la littérature psycholinguistique, le processus d'acquisition naturel mis en oeuvre par l'enfant reste mystérieux. Des modèles informatiques commencent toutefois à être proposés (Brent 1996). À notre avis, cette modélisation, et notamment celle des deux dernières phases, doit être symbolique et exacte. En effet, les erreurs de l'enfant lors de ces phases sont des erreurs de surgénéralisation, qui consistent à étendre un peu trop le domaine d'application des règles qu'il a inférées. Typiquement, un enfant dit d'abord « vous faîtes », la forme entendue, puis « vous faisez » avant de revenir à « vous faîtes ». La forme « faisez » prouve que des règles sont acquises. L'objet de cet article est de présenter certains modèles formels récents qui peuvent rendre compte de la phase III de cet apprentissage.

1.2. Acquisition de ressources syntaxiques

La linguistique informatique actuelle insiste sur la notion de ressource linguistique, c'est-à-dire sur les ensembles de données (liste de « mots vides », dictionnaire d'entités nommées, etc.) ou de programmes (lemmatiseur, étiqueteur grammatical, etc.) réunissant des informations linguistiquement pertinentes et utiles dans un cadre applicatif (par exemple dans un système de recherche d'information). La constitution d'une ressource est une tâche difficile et onéreuse : elle demande souvent des compétences complexes et toujours beaucoup de travail. Des techniques d'apprentissage automatique sont donc de plus en plus souvent utilisées pour acquérir de telles ressources à partir d'exemples ou de corpus, c'est-à-dire d'instances particulières de ce que les ressources produisent.

Les grammaires formelles sont des ressources particulières qui permettent non seulement d'évaluer la grammaticalité des énoncés, mais

aussi d'associer à ceux déclarés syntaxiquement corrects une structure. Cette structure, souvent un arbre étiqueté, rend compte de l'organisation interne de l'énoncé et est utile pour vérifier sa correction orthographique, ou pour le traduire. Les ressources grammaticales sont difficiles à acquérir, comme en témoignent les nombreuses grammaires incomplètes issues de projets ambitieux.

On comprend donc l'intérêt applicatif d'apprendre automatiquement une grammaire formelle à partir d'un corpus. On notera qu'il s'agit, comme pour l'enfant qui acquiert sa langue, d'apprentissage à partir d'exemples « positifs », c'est-à-dire d'énoncés syntaxiquement corrects. D'autres ressources peuvent concourir à la réalisation d'un tel objectif : une grammaire partielle écrite à la main, un corpus arboré d'une langue donnée.

1.3. Structure de l'article

Pour caractériser rigoureusement l'apprentissage automatique de grammaires à partir d'exemples, il faut définir les conditions dans lesquelles se déroule cet apprentissage. Celles-ci interviennent à différents niveaux. D'abord, il faut fixer l'ensemble des grammaires possibles, c'est-à-dire l'espace de recherche de l'algorithme d'inférence. Ensuite, il faut préciser la nature des données dont dispose l'algorithme. Enfin, et c'est ce qui fait défaut aux approches statistiques, il faut aussi définir un critère de succès de l'algorithme qui garantisse la correction du résultat. En inférence grammaticale, ces aspects constituent ce qu'on appelle un « modèle d'apprentissage ». Après avoir défini les grammaires utilisées, nous présenterons notre modèle d'apprentissage. Nous exposerons ensuite les résultats théoriques récemment obtenus ainsi que les implémentations pratiques en rapport. Ces travaux ont été effectués par les auteurs (avec d'autres chercheurs ayant aujourd'hui évolué dans d'autres directions), dans le cadre de l'Action de Recherche Coopérative de l'INRIA Gracq Acquisition de Grammaires Catégorielles (Lille, Nancy, Nantes, Rennes).

2. Formalismes syntaxiques

Dans cette partie, nous présentons différentes familles de grammaires qui serviront de « classe cible » à notre apprentissage, en insistant sur les propriétés qui les prédisposent à être automatiquement apprises.

2.1. Grammaires catégorielles AB

Les grammaires catégorielles sont issues d'une tradition logico-philosophique qui remonte à Aristote en passant par Husserl et Ajdukiewicz. Ce dernier s'en servit pour décrire les formules logiques bien formées ; leur

utilisation pour la formalisation de la syntaxe des langues est due à Bar-Hillel dans les années 50. A la différence des grammaires syntagmatiques introduites par Chomsky dans les mêmes années, las grammaires catégorielles sont :

- LEXICALISÉES. La grammaire d'une langue se réduit à un lexique, lequel associe à un mot une catégorie qui décrit son comportement syntaxique. Les règles, en nombre fini, sont indépendantes de la langue à décrire.
- APPLICATIVES / FONCTIONNELLES. Les règles qui assemblent deux catégories font un usage systématique de la notion « d'application fonctionnelle », telle qu'on la trouve dans la programmation fonctionnelle, en particulier typée (Scheme, Common Lisp, Ocaml, etc.) ou dans la logique intuitionniste : les catégories syntaxiques expriment l'interaction potentielle avec d'autres catégories syntaxiques par la possibilité d'appliquer une fonction à ses arguments.

LE PREMIER INGRÉDIENT D'UNE GRAMMAIRE CATÉGORIELLE EST L'ENSEMBLE CAT DE SES CATÉGORIES SYNTAXIQUES POSSIBLES. On se donne un ensemble fini B de catégories de base, dont S la catégorie des phrases, et souvent SN la catégorie des syntagmes nominaux et N la catégorie des noms communs. Cette liste finie ne varie pas d'une langue à une autre : il suffit de considérer toutes les catégories de base de toutes les langues humaines. Cat se construit en appliquant librement aux catégories de base et à celles déjà construites deux constructeurs : / (sur) et \ (sous). Ainsi, pour toutes catégories X et Y dans Cat, X/Y et Y\X sont aussi dans Cat. Cette notation à l'aide de symboles de fraction orientés remonte à Bar-Hillel (Ajdukiewicz n'utilisait que des fractions non orientées).

LE SECOND INGRÉDIENT DE LA GRAMMAIRE EST LE LEXIQUE qui donne pour chaque mot m la liste finie Lex(m) de ses catégories syntaxiques possibles : Lex(m) est donc un sous-ensemble fini de Cat.

LE TROISIÈME INGRÉDIENT EST L'ENSEMBLE DES RÈGLES universelles considérées : elles ne dépendent pas de la langue décrite mais de la variété de grammaires catégorielles considérée. Ces grammaires procèdent par catégorisation (un informaticien dirait par typage) de suites de mots qui sont souvent, mais pas toujours, des constituants. Une suite constituée d'un seul mot m admet pour catégories celles que le lexique assigne à m, c'est-à-dire Lex(m). Evidemment, comme pour un mot seul, une suite de mots peut appartenir à plusieurs catégories syntaxiques. Les règles universelles spécifient quelle peut être la catégorie syntaxique de la juxtaposition de deux suites de mots S1 et S2 en fonction des catégories syntaxiques possibles de S1 et S2. Pour les grammaires AB, les plus simples, il n'y a que deux règles :

- FA Si S1 est notamment de catégorie Y/X et que S2 est notamment de catégorie X alors S1 S2 est notamment de catégorie Y.

- BA Si S1 est notamment de catégorie X et que S2 est notamment de catégorie X\Y alors S1 S2 est notamment de catégorie Y.

La règle FA montre que la catégorie Y/X se comporte comme un « foncteur » qui s'applique à un argument X situé après, d'où son nom FA pour « Forward Application ». Symétriquement, la règle BA montre que la catégorie X\Y est un « foncteur » qui s'applique à un argument X situé avant et son nom signifie « Backward Application ».

Une suite de mots est une phrase produite par une grammaire AB s'il est possible de choisir, pour chaque mot m, une catégorie syntaxique dans Lex(m), de sorte que la suite de catégories ainsi obtenue se réduise en S au moyen des règles FA et BA. Les langages produits pas les grammaires AB sont exactement ceux que produisent les grammaires algébriques, aussi dites non contextuelles ou hors contexte. Les règles étant fixées et universelles, « apprendre » une telle grammaire revient donc à apprendre Lex(m) pour chaque mot m. C'est plus facile que d'essayer d'identifier un ensemble de règles syntagmatiques abstraites, sans lien avec le vocabulaire.

Exemple 1:

Soit B={S, N, SN} où SN désigne la catégorie des « Syntagmes Nominaux » et N la catégorie des « noms communs » et soit G la grammaire définie par le lexique suivant : Jean : SN, boit : (SN\S)/SN, SN\S, voit : SN\S, un : SN/N, homme : N, verre : N.

« Jean boit » est reconnu comme une phrase en choisissant SN\S comme catégorie pour « boit » : SN, SN\S –(BA)—> S et « Jean boit un verre » est reconnu comme une phrase en choisissant la catégorie (SN\S)/SN pour « boit » :

SN, (SN\S)/SN, SN/N, N –(FA)—> SN, (SN\S)/SN, SN
–(FA)—> SN, SN\S –(BA)—> S

Par contre, « verre boit » n'est pas reconnu comme une phrase car quelle que soit la catégorie que l'on choisisse pour « boit », ni FA ni BA ne s'applique à la suite de catégories : N, (SN\S)/SN ou N, SN\S.

Les grammaires qui n'assignent qu'une catégorie à chaque mot sont dit « RIGIDES » et jouent un rôle particulier pour l'apprentissage. Une grammaire peut être rendue rigide en considérant comme des mots différents les différents usages grammaticaux d'un même mot, tels « le » pronom et « le » article. Les grammaires AB rigides engendrent des langages algébriques non rationnels (non réguliers), mais pas tous les langages algébriques ni même tous les langages rationnels (cette classe est difficile à caractériser). La grammaire de l'Exemple 1 n'est rigide que si on considère qu'il y a deux mots « boit » ou si on supprime l'une des deux catégories de ce

mot.

2.2. Autres formalismes syntaxiques

Il existe de nombreuses variantes et extensions de ce formalisme de base. Une des plus simples propose d'ajouter de nouveaux schémas de règles qui étendent FA et BA (comme par exemple Z/Y Y/X—> Z/X). On définit ainsi les grammaires catégorielles dites combinatoires (Steedman 87).

Le calcul de Lambek (1958) est une autre extension, fondée sur la logique, des grammaires AB. Pour le définir, on ajoute à FA et BA deux nouvelles règles, qui sont un peu leur réciproque :

- \i si une suite S1 S2 est de catégorie Y avec S1 de catégorie X, alors S2 est de catégorie X\Y,
- /i si une suite S1 S2 est de catégorie Y avec S2 de catégorie X alors S1 est de catégorie Y/X.

Le calcul de Lambek est une logique avec deux implications : / et \, et quatre règles :FA, BA, \i, /i. Les catégories syntaxiques associées aux mots sont considérées, du point de vue logique, comme les hypothèses à partir desquelles il faut prouver S pour que la suite de mots soit une phrase.

Si, au lieu de suites d'hypothèses, on utilise des arbres d'hypothèses, on obtient le calcul de Lambek non associatif, à la base d'une variante multimodale très utilisée en linguistique.

Toutes les logiques ainsi obtenues sont dites sensibles aux ressources, parce qu'elles prennent en compte le nombre d'occurrences, la place voire la structure arborescente des hypothèses : en effet, pour une analyse linguistique, il est important de n'utiliser qu'une fois chaque mot ou syntagme, et leur place, voire leur situation structurelle importent. Par exemple, un verbe attend un objet et un seul : « J'appelle un chat » et « J'appelle un chat un chat » ne relèvent pas du tout de la même construction et « J'appelle un chat un chat un chat » est agrammatical.

Le formalisme récent des prégroupes (Lambek 1997) utilise un ensemble de catégories un peu différent et des règles de calcul proches des groupes ordonnés au sens mathématique.

Enfin, issues d'une autre tradition, les grammaires de dépendances définissent la syntaxe par des relations binaires orientées entre les mots plus que par la structure des constituants. Par exemple, un verbe transitif dans une phrase possède au moins deux dépendances : l'une vers la tête de son sujet, l'autre vers la tête de son complément direct. Les grammaires catégorielles de dépendances ou CDG (Dikovsky 2004) en sont une variante lexicalisée qui étendent les grammaires AB avec quelques règles dans le style des grammaires combinatoires générales. Une analyse avec une telle grammaire permet de construire l'arbre de dépendances de la phrase. Les catégories

simples gèrent les dépendances qui ne peuvent se croiser, dites projectives et des valences polarisées gèrent celles qui peuvent croiser l'arbre des dépendances projectives. Ces grammaires produisent tous les langages algébriques plus des langages non algébriques, tout en restant analysables en temps polynomial en fonction du nombre de mots dans la phrase.

2.3. Liens syntaxe/sémantique d'après Montague

Les travaux sémantiques du logicien Richard Montague dans des années 70 (Montague 1974, Dowty, Peters & Walls 1981) ont fortement influencé linguistes et informaticiens. Outre sa conception de l'intensionnalité, la principale contribution de Montague est la mise en œuvre effective du principe de compositionnalité qui stipule que « le sens d'une proposition ne dépend que du sens de ses constituants et de leur mode de combinaison syntaxique » (Partee 1990). Évoquons ici ce qui, dans ces travaux, peut avoir des incidences sur l'apprentissage de la syntaxe.

La représentation sémantique de toute unité syntaxique est, selon Montague, un terme fonctionnel typé. Le « typage sémantique » qu'il a défini, largement repris depuis, est construit à partir de deux types de base : le type t des valeurs de vérité et le type e des entités. Pour tous types a et b, a —>b est lui-même un type. a—>b est associé aux termes fonctionnels qui attendent une expression de type a pour donner une expression de type b. Par exemple, « Jean » désigne un individu de type e, tandis qu'un nom commun comme « homme » se comporte sémantiquement comme une fonction de type e—>t : appliquée à un individu, elle dit s'il est ou non un homme. On peut comprendre ces types en se référant à la théorie naïve des ensembles : les termes de type e sont les éléments d'un ensemble et t={0,1}. Un prédicat à une place P peut-être vu soit comme un sous-ensemble |P| d'entités soit comme une fonction caractéristique de type e—>t, qui envoie les éléments de |P| sur 1 et les autres sur 0. Similairement, un prédicat à deux places, comme « aimer », peut-être vu soit comme une partie des couples d'entités soit comme une fonction de type e—>(e—>t) qui, appliquée successivement à deux individus de type e, disons « Jean » et « Marie », vaudra 1 si Marie aime Jean et 0 sinon. Les termes de type t comme aime (Marie,Jean) traduisent des propositions complètes (en l'occurrence : « Marie aime Jean »).

En ce qui nous concerne ici, la révolution montagovienne en sémantique formelle consiste à automatiser la correspondance entre l'analyse syntaxique d'une proposition, telle que la réalise une grammaire catégorielle, et la construction de sa représentation sémantique logique. Ce parallèle commence par une traduction des catégories syntaxiques utilisées dans le lexique en types logiques, à l'aide d'un morphisme h tel que : h(SN)=e (un

syntagme nominal est une entité), h(N)=e—>t (un nom commun est un prédicat à une place, une propriété) et h(S)=t (une phrase est une proposition logique). Ce morphisme s'étend à toutes les autres catégories via la propriété suivante : pour toutes catégories X et Y, h(Y/X)=h(X\Y)=h(X)—>h(Y). On relie ainsi les notions de fonction, d'argument et de résultat, présentes à la fois dans les catégories et dans les types. Par exemple « aimer », qui a pour catégorie syntaxique (SN\S)/SN, est sémantiquement de type h((SN\S)/SN) qui vaut : h((SN\S)/SN)=h(SN)—>h(SN\S)=e—>(h(SN)—>t)=e—>(e—>t). Le lexique doit maintenant ainsi associer à chaque mot, en plus de sa catégorie syntaxique C, un terme de type h(C) qui représente sa sémantique.

Le parallèle établi par Montague fonctionne règle à règle. Chaque règle utilisée par l'analyse syntaxique d'une phrase par une grammaire catégorielle correspond à une étape du calcul de sa représentation sémantique globale à partir des traductions sémantiques des mots. Par exemple, la règle FA, qui applique la catégorie syntaxique Y/X de S1 à son argument X, associé à S2, se traduit sémantiquement par l'application d'un terme de type h(X)—>h(Y) (la sémantique de S1) à un terme de type h(X) (la sémantique de S2) pour donner un terme de type h(Y) (la sémantique de S1 S2). La sémantique d'une phrase est donc bien une proposition de type h(S)=t.

Nous verrons en partie 5.2 comment cette formalisation des liens entre syntaxe et sémantique peut être exploitée dans le cadre d'un modèle d'acquisition de la syntaxe où on suppose avoir en outre accès à des informations sémantiques.

3. Modèles de l'acquisition

Maintenant qu'ont été évoquées les grammaires à apprendre, c'est-à-dire la cible de l'apprentissage, restent à préciser les conditions de cet apprentissage. Nous présentons ici le modèle de Gold (1967), dans lequel se situe l'ensemble des travaux ici rapportés.

3.1. Modèle de Gold

En 1967, E. M. Gold propose (Gold 1967) un modèle formel de l'apprentissage de sa langue maternelle par un enfant, vu comme un problème d'inférence grammaticale à partir d'exemples positifs.

Dans ce modèle, l'apprentissage est un processus non borné pendant lequel l'apprenant, modélisé par un algorithme d'apprentissage A, est exposé à une suite infinie E d'exemples a_0, a_1, ..., a_i ,... appartenant au langage L, engendré par la grammaire cible G. On dit que E est une énumération du langage L si tout élément de E est un élément de L et si tous les éléments du langage L apparaissent tôt ou tard dans E. L'apprenant doit « deviner » à

partir de E quelle est la grammaire qui engendre ce langage L. À chaque nouvel exemple a_i, il propose une nouvelle grammaire hypothèse G_i. Il ne sait jamais si son hypothèse courante G_i est la bonne : à tout moment, un nouvel exemple peut l'amener à changer d'hypothèse.

On dira qu'il y a convergence si, au bout d'un nombre fini bien qu'inconnu d'exemples, l'hypothèse G_i de l'apprenant n'est plus modifiée et vaut constamment G+. Si cette hypothèse pérenne G+ engendre exactement le langage L, alors l'apprentissage est un succès. Gold dit qu'une classe de grammaires **G** est APPRENABLE À LA LIMITE s'il existe un algorithme A tel que pour tout langage L engendré par une grammaire G dans **G,** et pour toute énumération E de L fournie à A, l'algorithme A converge à la limite sur une grammaire G+ de **G** (éventuellement différente de G) qui engendre L.

La séquence d'exemples E a des propiétés notables :

- Elle ne contient que des exemples positifs, c'est-à-dire des éléments du langage ; elle ne dispose donc d'aucun contre-exemples, aussi appelés exemple négatif : c'est conforme aux conditions de l'apprentissage des enfants, mais cela rend la convergence difficile.
- La suite d'exemples est supposée ne comporter aucune erreur.
- Tous les objets du langage doivent obligatoirement apparaître dans la suite d'exemples. Ceci est possible y compris pour les langages infinis, parce que la séquence d'exemples est infinie
- Les exemples peuvent apparaître dans un ordre quelconque dans la séquence, et éventuellement plusieurs fois (ce qui permet d'énumérer indéfiniment un langage fini). Le modèle n'impose pas que l'algorithme donne le même résultat pour deux séquences différentes énumérant un même langage.

Dans ce modèle, la convergence d'un algorithme n'est pertinente que pour un ensemble de grammaires, et non pour une seule grammaire. En effet, ce dernier cas est trivial : il existe toujours un algorithme capable d'apprendre une grammaire donnée (puisqu'il suffit qu'il soit écrit pour renvoyer toujours cette grammaire). De même, un enfant naît avec la capacité d'apprendre n'importe quelle langue humaine, et pas uniquement sa langue maternelle. Intuitivement, plus la classe de grammaires à apprendre est vaste, plus il est difficile d'identifier une grammaire particulière de cette classe.

Soulignons enfin que la nature des exemples auxquels le sujet est exposé est fondamentale dans les résultats d'apprenabilité. La notion de « langage » employée dans la définition du modèle spécifie à la fois la nature des exemples et le critère de succès de l'apprentissage. Bien que ce modèle ait été introduit pour des grammaires produisant des suites de mots, il fonctionne aussi pour des grammaires dont les langages sont des ensembles d'objets plus structurés : langages d'arbres, de graphes, etc. L'important est

qu'un mécanisme génératif, typiquement une grammaire formelle, décrive finiment le procédé qui engendre tous les objets du langage. Comme nous le verrons, une classe de grammaires peut ne pas être apprenable à partir d'exemples qui sont de simples suites de mots, mais le devenir si ce sont des structures plus riches. Si les exemples sont des suites de mots, des phrases, on parlera d'APPRENABILITÉ LINÉAIRE ; si les exemples sont des arbres étiquetés dont les feuilles sont des mots, on parlera d'APPRENABILITÉ STRUCTURELLE.

3.2. Conditions suffisantes d'apprenabilité et de non apprenabilité

Il n'est pas toujours aisé d'établir ou de réfuter directement l'apprenabilité d'une classe de grammaires donnée au sens de Gold. Néanmoins il existe des critères garantissant l'une ou l'autre de ces possibilités.

3.2.1. Critères de non apprenabilité

Comme Gold l'a établi dans son article original, l'existence d'un « point limite » (nom malheureux puisqu'il s'agit un langage) dans une classe de langages suffit pour que toute classe de grammaires l'engendrant ne soit pas apprenable. Un point limite L d'une classe de langages est un langage qui est l'union d'une famille infinie de langages L_i de cette classe, chaque L_i contenant le précédent. On notera que si une classe de langages contient un point limite, alors il en est de même de toute classe plus grande.

À titre d'exemple, montrons que la classe entière des langages réguliers, et donc toute classe plus vaste, comme celle des langages algébriques, admet un point limite. Pour cette construction, un seul mot m suffit : les phrases seront donc des répétitions de i fois cet unique mot. Prenons pour L_i l'ensemble des phrases d'au plus i mots m répétés : L_i={m, mm,...,m^i}. Chaque L_{i+1} contient son prédécesseur L_i et l'union de tous les L_i est un langage régulier : $L=U_iL_i=m^*$ (l'ensemble de toutes les phrases possibles ne contenant que m). La classe a un point limite, aucune classe de grammaires l'engendrant n'est donc pas apprenable. On peut se convaincre du lien entre le point limite et la non apprenabilité comme suit : un algorithme auquel est fourni un nombre fini de répétitions de m de longueur maximale p ne pourra jamais deviner correctement s'il provient d'un langage L_i avec $i \geq p$ ou du langage limite L.

Le fait de démontrer qu'une classe de grammaires est non apprenable est une propriété très forte : elle signifie qu'il n'existe aucun algorithme, aussi astucieux soit-il, capable d'identifier toutes les grammaires de cette classe. On peut rapprocher cette notion de celle de « calculabilité » en informatique.

Le modèle de Gold fixe en quelque sorte les « conditions de possibilité » de l'apprentissage linguistique.

3.2.2. Critères d'apprenabilité

Donnons maintenant une propriété qui garantit l'apprenabilité : une classe de langages a une « élasticité finie » s'il est impossible de trouver une suite infinie de langages L_i dans cette classe ainsi qu'une suite infinie de phrases a_i avec $a_1,\ldots a_i$ dans L_i et a_{i+1} en dehors de L_i. L'existence d'un point limite interdit à une classe d'être finiment élastique : il suffit de considérer la suite infinie des mots répétant i fois m et la suite des langages L_i, sans même considérer le point limite L.

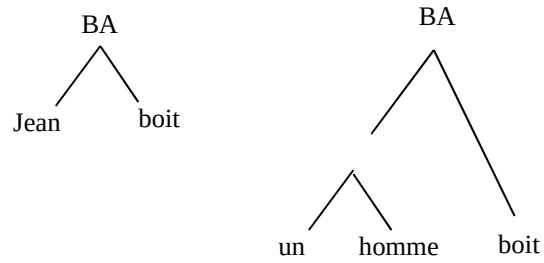
Toute classe de grammaires produisant une classe de langages d'élasticité finie est apprenable. Cette condition suffisante (mais pas nécessaire) d'apprenabilité est très utile : elle se laisse démontrer assez facilement, et, lorsqu'une classe de langages C est finiment élastique, il en est de même de toute classe C' manipulant des objets dans une relation simple avec ceux manipulés par C (Kanazawa 1998).

Dans (Shinohara 1991), l'élasticité finie, initialement définie pour des classes de langages, a été adaptée aux grammaires qui engendrent les langages de cette classe, donnant lieu au critère de « densité finie bornée ». Mentionnons que toute classe ne contenant que des langages finis satisfait ce critère et est donc apprenable.

4. Algorithme RG et ses extensions

Toute classe de grammaires produisant tous les langages algébriques n'est pas apprenable, et c'est exactement le cas de la classe des grammaires AB (aux phrase vides près) : elle n'est donc pas apprenable. Mais Kanazawa a néanmoins montré en 1994 que certaines sous-classes de grammaires AB sont apprenables en montrant la convergence de l'algorithme RG de (Buszkowski et Penn 1990). L'algorithme RG (ainsi nommé pour Rigid Grammars) et la preuve de sa correction sont essentiels à notre étude et méritent une présentation détaillée.

La cible de l'apprentissage est la classe des grammaires catégorielles AB rigides qui, rappelons-le associent à chaque mot une seule catégorie syntaxique. Les exemples de phrases fournies à l'algorithme sont des FA-structures. Une FA-structure est l'arbre de dérivation d'une phrase dont on a omis les catégories syntaxiques : les nœuds internes sont étiquetés du nom des règles (FA ou BA), et les feuilles des mots utilisés (leur lecture de gauche à droite donne la phrase). La figure suivante donne quelques FA-structures produites par la grammaire de l'Exemple 1 :



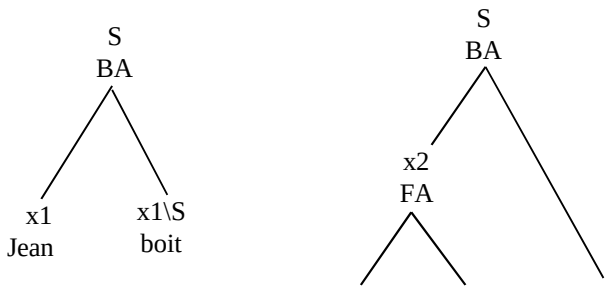
L'algorithme RG comporte trois étapes :

- L'étiquetage des FA-structures données comme exemples ;
- La construction d'une grammaire AB générant les exemples ; ce qui est toujours possible via l'étiquetage précédent
- L'unification des catégories syntaxiques associées à un même mot par cette grammaire, pour obtenir une grammaire rigide qui engendre toujours ces exemples, si toutefois c'est possible (et ce le sera lorsque les exemples énumèrent les FA-structures produites par une grammaire AB rigide).

La PREMIÈRE ÉTAPE consiste à étiqueter noeuds internes et feuilles des FA-structures de manière à obtenir un arbre de dérivation complet :

- Tout exemple étant une phrase de la grammaire à découvrir, la racine de chaque FA-structure doit recevoir l'étiquette S.
- On procède ensuite de la racine vers les feuilles. Supposons qu'un nœud interne ait l'étiquette N ; donnons, à l'aide d'une nouvelle catégorie x, des étiquettes à ses fils :
  - Si ce nœud est un FA alors le fils gauche reçoit l'étiquette N/x et le fils droit l'étiquette x.
  - Si ce nœud est un BA alors le fils gauche reçoit l'étiquette x et le fils droit l'étiquette x\N.

En suivant les règles de cette étape, les FA-structures de notre exemple sont étiquetées somme suit :



x2/x3	x3	x2\S
un	homme	boit

Cet étiquetage fabrique des arbres de dérivation complets pour chacun des éléments de l'ensemble D des exemples présentés et associe à chaque occurrence de mot dans une feuille de l'un de ces exemples une catégorie.

La DEUXIÈME ÉTAPE consiste à définir GF(D) (pour « General Form »), la grammaire AB qui associe à chaque mot toutes les catégories qu'ont reçues ses diverses occurrences dans l'étape précédente. Dès qu'un mot figure plusieurs fois dans un même exemple ou dans des exemples différents, il reçoit plusieurs catégories syntaxiques et GF(D) n'est pas rigide. Dans notre exemple, GF(D) vaut : Jean : x1, boit : x1\S et x2\S, un : x2/x3, homme : x3. Comme l'étiquetage le montre, GF(D) engendre toutes les FA-structures de D. Mais si l'algorithme s'en tenait là, le processus ne convergerait pas puisque chaque nouvel exemple oblige à introduire de nouvelles catégories. On a du reste vu que les grammaires AB sans restriction ne sont pas apprenables, quelle que soit la méthode.

C'est la raison d'être de la TROISIÈME ÉTAPE qui unifie les différentes catégories associées à un même mot dans GF(D). Cette opération consiste à substituer une catégorie à certaines autres, élémentaires ou non, dans GF(D), de sorte que chaque mot n'ait plus qu'une seule catégorie, et que la grammaire soit donc rigide. Dans notre exemple, il suffit de substituer une nouvelle variable x4 à x1 et x2 (pour unifier les catégories de « boit »). La grammaire résultante vaut alors désormais : Jean : x4, boit : x4\S, un : x4/x3, homme : x3. C'est cette grammaire qui est proposée comme hypothèse par l'algorithme d'apprentissage auquel on a fourni D. Si les éléments de D ne font pas partie d'un même langage de FA-structure, cette étape peut échouer car tout ensemble fini de catégories n'est pas nécessairement unifiable.

Kanazawa (1994, 1998) a montré que cet algorithme converge : si les FA-structures sont produites par une grammaire AB rigide G, alors on peut toujours unifier les catégories de GF(D) et, au bout d'un certain temps, la grammaire hypothèse vaudra toujours une grammaire G+, telle que le langage de FA-structures de G+ est le même que celui de G.

La construction d'une grammaire rigide hypothèse à partir d'un ensemble de FA-structures D est quadratique en la taille de D, ce qui est relativement efficace. De plus, l'algorithme RG peut-être appliqué incrémentalement en partant de la grammaire hypothèse obtenue à l'étape précédente et du nouvel exemple, sans recalculer GF(D) à chaque fois. Par contre le nombre d'exemples nécessaires, dans le pire des cas, à la construction de la grammaire cible, est inconnu ; on peut montrer qu'il est exponentiel en fonction de la taille de la grammaire cible, qui est en général initialement inconnue.

Le principal défaut de RG est qu'il nécessite des FA-structures : n'est-ce pas présupposer trop d'informations, et d'un point de vue pratique où les trouver ? Nous verrons en partie 5.2 qu'on peut interpréter les FA-structures comme résultant d'informations sémantiques. L'enfant y a ainsi peut-être accès, même si elles ne figurent encore dans aucun corpus.

Kanazawa a étendu ce premier résultat, fondamental, d'une part en enrichissant la classe de grammaires catégorielles considérée, et d'autre part en réduisant l'information structurelle véhiculée par les exemples. Par exemple, les grammaires AB rigides sont aussi apprenables à partir de simples phrases (suites de mots sans FA-structure). Toutefois, dans ce cas, l'algorithme d'apprentissage doit être modifié : une étape préliminaire doit être ajoutée, consistant à essayer toutes les FA-structures possibles qui « collent » aux phrases fournies. Ce calcul peut excéder la capacité des ordinateurs car il y en a un nombre exponentiel par rapport à la taille des phrases.

De même, pour tout nombre entier k>1, la classe des grammaires catégorielles qui associent au plus k catégories différentes à un même mot (appelées aussi « k-valuées ») est apprenable à partir de FA-structures et même à partir de phrases, mais au prix de calculs encore plus considérables. On prendra garde à l'apparente contradiction suivante : toute grammaire AB étant k valuée pour un certain k, pourquoi la classe entière des grammaires AB n'est-elle pas apprenable, contrairement à ce que nous disions précédemment ? La réponse est que l'algorithme d'apprentissage ne peut être défini que si la borne k est connue à l'avance.

5. Résultats et expériences

Les fondements du domaine étant posés, nous pouvons maintenant évoquer les résultats récents auxquels ont contribué les auteurs de cet article. Nous commencerons par les résultats de non apprenabilité, qui permettent à la fois de cartographier les grammaires catégorielles du point de vue de l'apprentissage et de conforter certaines hypothèses linguistiques. Nous poursuivrons par les résultats d'apprenabilité qui, outre leur intérêt théorique, permettent d'envisager une utilisation des algorithmes d'apprentissage pour construire automatiquement des grammaires électroniques.

5.1. Apprenabilité et non apprenabilité linéaire

5.1.1. Non apprenabilité linéaire

Les résultats de non apprenabilité que nous avons obtenus concernent surtout l'apprentissage à partir de phrases, où les exemples sont de simples

suites de mots. Cela conforte l'hypothèse selon laquelle un minimum de structure doit être fourni à l'enfant, notamment par la prosodie ou la sémantique.

Contrairement aux grammaires catégorielles AB rigides, les grammaires de Lambek rigides, n'associant qu'une seule catégorie à chaque mot et utilisant les 4 règles évoquées en 2.1 et 2.2, ne sont pas linéairement apprenables (Foret & LeNir 2002). Les variantes les plus courantes (grammaires de Lambek non-associatives, autorisant la séquence vide...) ne sont pas non plus linéairement apprenables (Béchet & Foret 2003b). Finalement, même les grammaires de prégroupes, pourtant plus simples et plus adaptées à décrire des phrases, ne sont pas non plus linéairement apprenables (Béchet & Foret 2003c). Dans chaque cas, la preuve consiste à exhiber un « point limite » (cf. partie 3.2.1) pour la classe de langages considérée. Comme les langages k-valués (au plus k catégories par mot) contiennent les langages rigides, ces résultats de non-apprenabilité s'étendent aux grammaires de Lambek k-valuées.

Donnons un exemple de construction de point limite pour les grammaires de Lambek rigides avec séquence vide. Dans ce cas, la suite de langages L_i est caractérisée par une suite de grammaires G_i définies sur un vocabulaire de trois mots, comme suit : a :P/P, b :Q/Q et c : D_i où D_i est défini par D_1=S, D_{i+1}=(D_i/(P/P))/(Q/Q). Le langage L_i engendré par G_i est c $(b*a*)^i$ (le mot c suivi de au plus i successions de groupes de mots de la forme b...ba...a). Le point limite, c{a,b}*, est aussi produit par une grammaire de Lambek rigide : a : P/P, b : P/P, c : S/(P/P).

Dans certains formalismes, ce qui pose problème pour l'apprentissage à partir de phrases, c'est l'existence d'arguments optionnels. Ceci est un problème pour toutes les classes de grammaires. La classe des CDG rigide n'échappe pas non plus à cette règle. La non apprenabilité linéaire de ces formalismes signifie qu'ils ont un pouvoir d'expression élevé, même avec une seule catégorie par mot et que des structures sont nécessaires. Elle indique des impasses à éviter.

5.1.2. Apprenabilité linéaire

Pour obtenir des résultats d'apprenabilité linéaire sur des sous-classes de grammaires de Lambek, il faut imposer des contraintes supplémentaires sur les catégories. L'arité d'une catégorie est le nombre d'arguments avec lesquels elle peut se combiner. Si on borne l'arité des catégories tout en limitant le nombre de catégories pouvant être associées à chaque mot, alors l'apprentissage des grammaires de Lambek non associatives à partir de simples phrases devient possible (Béchet & Foret 2003a). Cela fait néanmoins beaucoup de contraintes.

Quelques variantes et extensions des grammaires catégorielles AB sont aussi linéairement apprenables : citons les grammaires de liens rigides (Béchet 2003) et les grammaires catégorielles de dépendance rigides sans arguments optionnels (Béchet Dikovsky, Foret & Moreau 2004). En fait, hormis les grammaires AB k-valuées, il y a peu de classes de grammaires linéairement apprenables. Et même pour ces classes, l'algorithme d'apprentissage est très peu efficace (il teste toutes les structures possibles pouvant être associées à chaque phrase).

5.1.3. Apprentissage à partir d'une grammaire noyau

Est-il possible de mettre en oeuvre un algorithme de type RG pour acquérir en temps raisonnable une grammaire à partir de données réelles, sans disposer de FA-structures ? Une piste, explorée dans (Moreau 2004a, 2004b), consiste à fournir à l'algorithme d'apprentissage une partie de la grammaire à apprendre. En effet, de nombreuses grammaires partielles ont déjà été écrites « à la main ». Mieux vaut essayer de partir de ces ressources que de chercher à tout apprendre à partir de rien. Un algorithme d'apprentissage partiel a donc été proposé, dont le fonctionnement doit permettre de tirer profit des informations d'une grammaire initiale (le noyau) pour pallier l'absence des informations qui auraient été apportées par des structures.

L'intérêt de cette approche repose entièrement sur la grammaire initiale : elle doit être suffisamment complète pour qu'un nombre significatif de mots des phrases fournies en exemple soient connus. Dans ce cadre, on peut tirer profit de la loi de Zipf, utilisée notamment par l'étiqueteur de Brill. Celle-ci garantit que, sur l'ensemble des mots d'un texte, une faible proportion suffit à représenter une grande partie de ce texte (en nombre d'occurrences). Or, précisément, ce sont les mots les plus fréquents d'un langage qui sont le plus facile à répertorier et définir (mots grammaticaux tels que déterminants, pronoms, prépositions, conjonctions, etc.), soit manuellement soit par conversion de dictionnaires existants dans d'autres formalismes.

La faisabilité de cette approche a été évaluée à l'aide du lexique anglais proposé par les auteurs des grammaires de liens, Sleator et Temperley (1991). Même si ces données sont relativement simples et de taille limitée, il s'agit d'une grammaire assez complète de l'anglais. En ce sens, cette première expérimentation est représentative des problèmes que pose le passage à l'échelle de véritables données textuelles pour l'apprentissage automatique. Les résultats obtenus semblent indiquer que l'apprentissage partiel répond de façon satisfaisante au problème de la nature des données requises par l'algorithme d'apprentissage (Moreau 2004b).

5.2. Apprentissage à partir d'informations sémantiques

Le lien établi par Montague (cf. partie 2.3) entre syntaxe et sémantique peut être exploité dans un modèle d'acquisition de la syntaxe. Nous avons vu, en effet, que les enfants acquièrent des notions de sémantique lors de la phase II de leur apprentissage du langage, donc avant d'en maîtriser la syntaxe.

Plusieurs cas peuvent se présenter, suivant les hypothèses que l'on fait sur la nature des informations disponibles dans l'environnement d'apprentissage, et suivant ce qui a déjà été préalablement acquis. Une première hypothèse consiste à supposer que la sémantique lexicale (l'association entre un mot et le terme qui le traduit) est acquise en premier, et que les données d'apprentissage auxquelles est soumis l'apprenant sont des associations entre un énoncé syntaxiquement correct et une représentation de son sens. On peut alors montrer que, moyennant une version légèrement renforcée du Principe de Compositionnalité, l'ensemble de ces informations suffit à caractériser, quand elle existe, une unique FA-structure (Tellier 1999). On sélectionne ainsi une FA-structure pertinente (au sens où elle mènera à une bonne traduction logique) parmi l'ensemble (très nombreux) de toutes les FA-structures possibles. Dans ce cas, l'information sémantique permet d'expliquer d'où proviennent les données nécessaires à l'algorithme RG.

Mais il est aussi possible de s'affranchir complètement de la notion de FA-structure en faisant l'hypothèse que les exemples disponibles à l'apprenant sont des énoncés syntaxiquement corrects dans lesquels les mots sont étiquetées par leur type sémantique. Savoir qu'un mot correspond à un prédicat à un argument (de type e—>t) ou à un prédicat à deux arguments (de type e—>(e—>t)) etc. constitue en effet une hypothèse raisonnable. Or, nous l'avons vu, la correspondance entre syntaxe et sémantique chez Montague passe par l'existence d'un morphisme h qui transforme toute catégorie syntaxique C en un type logique h(C). L'apprentissage des catégories C est, on s'en doute, grandement facilité par la donnée des h(C).

L'apprenabilité de grammaires catégorielles à partir de phrases typées a été étudiée en détails dans (Dudau, Tellier & Tommasi 2001a, 2001b, 2003) et (Dudau & Tellier 2004a). Pour caractériser les classes de grammaires catégorielles apprenables par phrases typées, il faut faire intervenir le morphisme h. Les principaux résultats obtenus dans ce cadre sont les suivants : la classe des grammaires catégorielles AB (resp. de Lambek) telles que pour toute association d'un mot m et d'un type sémantique h(C), il existe une seule catégorie C dans Lex(m) appartenant à la grammaire cible (ce qui est en fait une autre forme de rigidité) est apprenable au sens de Gold par phrases typées.

Un nouvel algorithme d'apprentissage spécifique a été défini, implémenté et testé (Dudau-Sofronie & Tellier 2004b) dans ce contexte. Cet algorithme est fondamentalement différent de RG. RG procède par généralisations successives : la suite des grammaires hypothèses qu'il formule reconnaissent en effet des langages de plus en plus grands. L'apprentissage par phrases typées, lui, opère par restrictions progressives de l'espace de recherche : c'est un algorithme par spécialisation.

Le problème qui se posait pour mener à bien des expériences à partir de cet algorithme, c'est qu'aucun corpus étiqueté sémantiquement n'est pour l'instant disponible. Il a donc fallu en produire un. La stratégie employée a consisté à utiliser un outil d'étiquetage « part-of-speech » (le Tree-Tagger) et à remplacer les étiquettes obtenues par les types sémantiques correspondant. Dans la plupart des cas, en effet, l'étiquette « part-of-speech » détermine un type sémantique unique (c'est le cas des déterminants, de la plupart des noms communs et des adjectifs, etc.). Mais certaines ambiguités subsistaient : par exemple le Tree-Tagger ne distingue pas les verbes intransitifs des verbes transitifs, qui doivent recevoir des types différents. Dans ce cas, c'est une heuristique au niveau de la phrase qui permettait généralement de trancher et d'éviter plusieurs affectations possibles de types à chaque mot. Il est important de noter que l'information « part-of-speech » initiale s'effaçait complètement derrière l'information de type, et que l'apprentissage n'était pas biaisé par cette étape intermédiaire. Deux mots distincts associés à la même étiquette « part-of-speech » pouvaient très bien, après apprentissage, recevoir des catégories syntaxiques différentes, et deux mots distincts avec des étiquettes différentes se retrouver avec la même catégorie.

Le corpus employé était la transcription de contes écrits par des enfants pour des enfants, et corrigés par leur enseignant. Il était constitué de phrases simples et courtes. Les difficultés rencontrées dans ces expériences, comme dans l'apprentissage à partir de phrases, était la faible redondance du vocabulaire, et le grand nombre de grammaires compatibles obtenues.

5.3. Apprentissage à partir de structures partielles de dérivation

Si on dispose des structures de dérivation des phrases, il est plus facile d'apprendre les grammaires qui les produisent. Nous passons en revue ici divers résultats d'apprenabilités avec structures qui étendent ceux de Kanazawa.

5.3.1. Résultats d'apprenabilité

Tout d'abord, l'algorithme RG peut être adapté pour apprendre les grammaires de Lambek rigides à partir des dérivations respectant une certaine forme normale associées à chaque phrase (Bonato & Retoré 2000). Les grammaires de Lambek rigides se situent ainsi vraiment à la frontière de ce qui est apprenable : elles le sont à partir de (certaines) structures et à partir de phrases typées, mais pas à partir de phrases seules.

On peut aussi adapter RG pour qu'il se contente d'arbres binaires de mots, autrement dit de FA-structures dont les noms de règles (FA et BA) sont omis (Le Nir 2003). Cette perte d'informations doit être compensée par un temps de calcul accru de l'algorithme d'apprentissage, qui teste alors tous les étiquetages possibles, mais c'est un compromis entre l'absence de structures et l'utilisation de structures trop riches. Cet algorithme peut être appliqué aux grammaires de Lambek non associatives k-valuées de degré de réduction p -le degré de réduction borné par p permet de se ramener à des grammaires AB équivalentes en p étapes logiques.

Les grammaires combinatoires générales, avec en plus de FA et BA d'autres règles (cf. 2.2) mais sans règles d'introduction à la Lambek, n'avaient pas été couvertes par Kanazawa. En utilisant l'élasticité finie, (Moreau 2005, 2006) a montré que les grammaires plates k-valuées, les grammaires à arguments bornés et les grammaires par consommation stricte d'arguments k-valuées sont structurellement apprenables (voire linéairement apprenable sous certaines restrictions).

L'apprenabilité des grammaires catégorielles de dépendance (CDG) a aussi été étudiée. Si on utilise comme entrées des réseaux de dépendances (des arbres de dépendances appauvris qui jouent le même rôle pour ces grammaires que les FA-structures pour les grammaires AB), elles ont une densité finie bornée et sont donc apprenables à partir des réseaux de dépendance. Les CDG rigides sont aussi apprenables à la limite à partir de phrases (Béchet, Dikovsky, Foret Moreau, 2004).

Enfin, à l'instar des précédentes, les grammaires minimalistes de Stabler sont linguistiquement pertinentes et elles peuvent générer plus que les langages algébriques. Elles admettent une présentation qui ressemble à celle des grammaires de Lambek avec produit et sont apprenables à partir de structures de dérivations similaires (Bonato & Retoré 2000).

5.3.2. Apprentissage à partir d'un corpus arboré

Il existe depuis maintenant plusieurs années des corpus dits « arborés », qui contiennent les analyses syntaxiques complètes de phrases. Ces ressources sont-elles utilisables par les algorithmes d'apprentissage à

partir de structures dont nous venons de parler ? En fait, l'algorithme RG est difficilement applicable directement. D'une part, les grammaires résultats sont supposées être rigides, ce qui n'est pas suffisant pour les langues naturelles. D'autre part, les phrases en entrée de l'algorithme doivent être présentées sous la forme de FA-structures, qui n'est pas le format des corpus arborés.

L'apprentissage par exemples structurés a portant commencé à être testé en partant du corpus arboré de phrases françaises issues du journal « Le Monde » (d'une taille d'un million de mots) développé à l'Université Paris 7 (Abeille et al. 2003). La démarche suivie consitait d'abord à transformer le corpus d'arbres en « FA-structures annotées », puis à appliquer un algorithme d'apprentissage inspiré de RG sur ces structures (Poupard, Béchet & Foret 2006).

6. Critiques et problèmes

Nous revenons ici sur les avantages et inconvénients de l'approche défendue dans cet article.

6.1. Sur les résultats théoriques d'apprenabilité

Le critère d'identification à la limite, défini comme base de l'apprenabilité, fait du modèle de Gold une formalisation principalement théorique du processus d'apprentissage. En effet, si ce critère garantit bien l'existence d'un algorithme d'apprentissage pour les classes de grammaires apprenables, il ne dit rien de son efficacité. Ainsi, un algorithme qui se contenterait d'énumérer tous les éléments de la classe à apprendre jusqu'à en trouver un compatible avec les exemples serait tout à fait recevable dans le modèle de Gold, bien qu'inutilisable en pratique. La mise en oeuvre de l'apprentissage nécessite au minimum que l'algorithme soit capable de construire une grammaire à partir des données dont il dispose, si possible en un temps raisonnable.

De tels algorithmes d'apprentissage efficaces existent, bien sûr : c'est le cas de RG. Mais pour son application aux langues naturelles, il présente deux défauts majeurs : la limitation aux grammaires rigides et la nécessité de structures complexes en entrée. En théorie, nous n'avons pas toujours besoin des structures mais en pratique elles sont très utiles.

6.2. Sur les expérimentations pratiques

La mise en oeuvre d'algorithmes d'apprentissage de grammaires sur des données réelles se heurte donc aux problèmes classiques d'efficacité, de disponibilité de données utilisables et plus généralement de choix nécessaires au passage de la théorie à la pratique. Différents paramètres entrent en jeu :

- **Nature des données fournies à l'algorithme en entrée :** de toute évidence, plus l'algorithme dispose d'informations (structure de dérivation, étiquetage syntaxique et/ou sémantique, etc.), plus celui-ci peut apprendre correctement et/ou rapidement la grammaire. L'apprentissage à partir de phrases est extrêmement difficile à réaliser sur des données de grande taille, à cause de l'explosion combinatoire des possibilités. A l'inverse, l'apprentissage à partir de structures précises présente l'inconvénient de contraindre fortement le résultat : il faut alors que les données utilisées soient très fiables. Comme ce type d'information n'est pas souvent disponible, surtout dans le formalisme grammatical utilisé, cela implique éventuellement la conversion de données existantes, avec un risque de perte d'information.

- **Complexité algorithmique, complexité du processus d'apprentissage :** dans le contexte de l'apprentissage, la complexité algorithmique classique (le temps d'exécution de l'algorithme en fonction de la taille des données en entrée) ne suffit pas à caractériser l'efficacité du processus. En effet, selon la classe de grammaires cible et la méthode d'apprentissage utilisée, l'algorithme peut avoir besoin d'un nombre plus ou moins grand d'exemples pour converger. De plus, certaines propriétés de la méthode d'apprentissage améliorent significativement ses performances : on préférera toujours un algorithme incrémental, capable de synthétiser les informations des exemples déjà vus en évitant de les traiter de nouveau lorsqu'un nouvel exemple est proposé.

- **Niveau d'adéquation avec le modèle de Gold :** le modèle de Gold est trop strict ou trop idéalisé pour être utilisé tel quel en pratique. La convergence sur une séquence infinie d'exemples est impossible à tester, puisque seules des données finies sont disponibles. La pratique contraint souvent à produire une grammaire approchée, ou un nombre fini de grammaires, sans atteindre la convergence.

7. Conclusions

Reprenons pour finir nos résultats et expériences à la lumière des motivations premières exposées dans la présentation de l'article.

7.1. Formalisation de l'acquisition naturelle de la syntaxe

Le critère d'apprenabilité de Gold est un modèle raisonnable de l'apprentissage linguistique. Il rend bien compte de l'apprentissage à partir d'exemples positifs uniquement. Il est toutefois sujet à deux critiques. L'une, à laquelle répondent en partie nos travaux, concerne la nature des exemples disponibles. L'autre porte sur l'absence de contraintes de complexité algorithmique dans la définition de la convergence.

Sur la nature des exemples, nos résultats confortent ce qu'a mis à jour la psycholinguistique : avec de simples phrases, l'identification est impossible ou nécessite un temps de calcul démesuré. La psycholinguistique nous dit que la prosodie et la sémantique sont des ingrédients nécessaires au processus d'apprentissage. En raison de la difficulté de leur formalisation, elles n'ont pu être aussi bien représentées que la syntaxe dans nos modèles. Pourtant, les FA-structures (ou même les simples arbres binaires qui les supportent) sont en lien avec les structures prosodique et sémantique, et dès qu'elles sont présentes dans les exemples, des classes de grammaires linguistiquement intéressantes deviennent apprenables.

Surtout, l'ajout d'informations sémantiques agit en catalyseur sur le processus d'acquisition de la syntaxe. L'apprentissage avec l'indication élémentaire de la structure argumentale de chaque mot fonctionne, et en temps relativement raisonnable. Ce résultat montre non seulement la validité du modèle de Gold, mais aussi la pertinence des grammaires catégorielles, grâce à leur lien étroit avec la structure logique de la phrase, décrite à la Montague. De plus, l'algorithme d'apprentissage à partir de phrases typées, évoqué en partie 5.2., a un fonctionnement qui semble plus proche de l'apprentissage « naturel » que l'algorithme RG, parce qu'il opère par restrictions progressives de l'espace des grammaires possibles, plutôt que par généralisations successives. Les enfants procèdent de même.

Sur le point de l'efficacité, il est difficile de comparer les conditions d'apprentissage « naturelle » et « artificielle ». Beaucoup de paramètres nous échappent et font encore l'objet de discussions : Combien de phrases sont entendues par l'enfant durant la période de son apprentissage ? Quelle est leur longueur moyenne ? Et surtout : quelle est la nature de la grammaire acquise ? On rejoint ici une réflexion linguistique générale sur la forme possible des grammaires des langues humaines et leur part d'inné. Cette réflexion est elle-même alimentée par la complexité algorithmique de l'analyse et par celle de… l'apprentissage. L'analyse a conduit à la notion de grammaire légèrement contextuelles analysables en temps polynomial (comme les grammaires minimalistes, évoquées ici).

En matière d'apprentissage, Chomsky a promu récemment une approche dite *Principes et Paramètres*, selon laquelle la grammaire

s'acquiert à partir d'une matrice grammaticale innée, appelée grammaire universelle, en fixant un nombre fini de paramètres. Le puzzle de l'apprentissage se résoud alors de lui-même, et ne pose plus de difficultés algorithmiques. Mais bien peu de paramètres ont été réellement identifiés et ceux qu'on connaît concernent des points subtils d'interface entre syntaxe et sémantique ; il est donc difficile de faire reposer une théorie de l'apprentissage sur ces seuls indices. Mentionnons néanmoins que les modèles proposés ici intègrent le fait que la variation langagière réside dans le lexique. L'apprentissage de grammaires de types logiques distingue aussi l'universel (les règles) du particulier (les affectations de catégories).

7.2. Construction automatisée de grammaires : méthodes empiriques ou symboliques ?

Qu'apporte le processus d'identification à la limite de Gold à la construction automatisée de grammaires électroniques ? En supposant que les données nécessaires soient disponibles, des algorithmes comme les nôtres sont-ils utilisables, en particulier quant à leur temps d'exécution ? Où trouver les données pertinentes ? Comment les construire ?

Sur des données réelles, seuls sont viables les algorithmes qui évitent d'énumer la classe de grammaires à apprendre en construisant une grammaire hypothèse à partir des exemples en temps polynomial. Ainsi fonctionne RG, même si ses défauts sont maintenant bien identifiés : limitation aux grammaires rigides, nécessité de structures complexes en entrée, et choix arbitraire d'arrêter le processus (puisqu'on ne sait jamais si la convergence est atteinte).

Dans la communauté du traitement automatique des langues, la construction automatisée de grammaires électroniques est un thème crucial. D'autres travaux que les notres s'attaquent à l'acquisition de grammaires lexicalisées en utilisant des méthodes statistiques. Par exemple, Osborne & Briscoe (1997) utilisent une version stochastique des grammaires catégorielles (GCS), dans lesquelles les catégories associées à un mot sont affectées d'un coefficient de probabilité qui permet de classer les analyses syntaxiques des phrases en fonction de leur « pertinence » relative. Les auteurs proposent de calculer une estimation de ces coefficients en utilisant plusieurs méthodes d'apprentissage sur un corpus de phrases analysées. En fait, contrairement à ce qui est présenté ici, les catégories associées aux mots sont déjà présentes dans le corpus, seule leur probabilité est inconnue. Dans le même esprit, Zettlemoyer & Collins (2005) modélisent la syntaxe par des grammaires catégorielles combinatoires stochastiques. Les catégories possibles sont choisies parmi un ensemble fixé à priori. Collins (1996), lui,

utilise les grammaires de dépendances. Dans ce cas, l'apprentissage consiste à déterminer de manière probabiliste quelles sont les meilleures dépendances entre les mots d'une phrase. Les grammaires résultant de cet apprentissage sont stochastiques et assez proches des grammaires de liens, mais le lexique engendré n'est pas synthétique : les catégories associées à un mot sont engendrées séparément les unes des autres.

Hockenmaier (2003) a extrait une grammaire de l'anglais sous forme de grammaire catégorielle combinatoire (Steedman 1987) à partir du corpus arboré de la Penn Treebank. De manière assez similaire, Moot a extrait un lexique du Hollandais à partir du corpus CGN, dans lequel on dispose d'informations syntaxiques précises (Moot 2001, 2003, 2006). Dans les deux cas, la construction de la grammaire consiste essentiellement à convertir les données d'origine dans le formalisme syntaxique cible. Ceci nécessite que les différentes constructions syntaxiques possibles spécifiques de la langues soient auparavant étudiées au cas par cas, en grande partie manuellement. En ce sens, il s'agit plus d'extraction d'informations syntaxiques existantes que d'induction de grammaire. L'absence de cadre formel au processus d'apprentissage ne permet pas de garantir l'exactitude de la grammaire obtenue. Il semble ainsi difficile de démontrer l'absence de surgénéralisation par rapport au langage cible.

Enfin, une autre technique standard consiste à extraire d'une partie d'un corpus arboré des sous-arbres attachés à un mot ; ces sous-abres sont ensuite traduits en catégories syntaxiques de ce mot. Vue la taille des données, les erreurs d'assignation de catégorie sont très nombreuses : elles peuvent provenir des erreurs, en général nombreuses, du corpus arboré, mais aussi de la découpe du sous-arbre, voire de sa traduction en catégorie syntaxique. La grammaire résultat associe une centaine de catégories différentes à chaque mot et ne peut être utilisée « normalement », c'est-à-dire en essayant toutes ces catégories possibles. Les systèmes les plus performants font appel à des probabilités pour construire les séquences de catégories les plus probables sur l'ensemble de la phrase (c'est ce qu'on appelle le supertagging). D'autres fabriquent simultanément toutes les analyses possibles qu'ils représentent par des forêts partagées : la difficulté est alors d'extraire les bonnes analyses parmi les milliers engendrées.

Ce qui frappe dans ces travaux, c'est que les grammaires construites sont en grande partie fausses et que le seul critère de convergence est la couverture des hypothèses : la bonne analyse est-elle la plus probable, figure-t-elle dans les dix (ou vingt, cent etc.) analyses les plus probables ? Si ce n'est pas le cas, on augmente la proportion de corpus utilisée pour l'extraction de la grammaire, et on recommence. Malgré leurs défauts, les méthodes inspirées du paradigme de Gold peuvent contribuer à obtenir

automatiquement des grammaires plus fiables, avec des assignations de catégories moins souvent erronées.

Bibliographie

ABEILLÉ, Anne ; CLÉMENT, Lionel ; TOUSSENEL, François (2003), « Building a treebank for French ». Dans ABEILLÉ, A. (ed.), *Treebanks: Building and Using Parsed Corpora*.p. 165-188, Dordrecht : Kluwer.

BÉCHET, Denis (2003), « k-valued link grammars are learnable from strings », Proceedings of the 8th Conference on Formal Grammars, Vienna, Austria, p. 9-18.

BÉCHET, Denis ; DIKOVSKY, Alexander ; FORET, Annie (2005), « Dependency Structure Grammars », Proceedings of the 5th LACL, LNAI 3492, Springer-Verlag, p. 18-34.

BÉCHET, Denis ; DIKOVSKY, Alexandre ; FORET, Annie ; MOREAU, Erwan (2004), « On learning discontinuous dependencies from positive data », Proceedings of the 9th Conference on Formal Grammars, Nancy, France, p. 1-16.

BÉCHET, Denis ; FORET, Annie (2003a), « Apprentissage de grammaires de Lambek rigides et d'arité bornée pour le traitement automatique des langues », actes de CAP' 2003, Presses universitaires de Grenoble, p. 155-167.

BÉCHET, Denis ; FORET, Annie (2003b), « k-valued non-associative Lambek categorial grammars are not learnable from strings », Proceedings of the 41st ACL, Sapporo, Japan, p. 351-358.

BÉCHET, Denis ; FORET, Annie (2003c), « Remarques et perspectives sur les langages de prégroupe d'ordre ½ », actes de TALN, Batz sur Mer, France, p. 309-314.

BÉCHET, Denis ; FORET, Annie (2006), « k-valued non-associative Lambek grammars are learnable from generalized functor-argument structures », *Theoretical Computer Science*, 355-2: p. 139-152.

BÉCHET, Denis ; FORET, Annie ; TELLIER, Isabelle (2004), « Learnability of pregroup grammars », Proceedings of ICGI, LNAI 3264, Springer-Verlag, p. 65-76.

BONATO, Roberto ; RETORÉ, Christian (2000), « Learning Rigid Lambek Grammars and Minimalist Grammars from Structured Sentences », Proceedings of the 3rd Workshop on LLL, Strasbourg, 2001.

BOYSSON-BARDIE Bénédicte (1996), *Comment la parole vient aux enfants*, Odile Jacob, Paris.

BRENT, Michael R. (1996), *Computational approaches to language acquisition*, MIT Press.

BUSZKOWSKI, Wojciech ; PENN, Gerald (1990), « Categorial Grammars determined from linguistic data by unification », *Studia Logica* 49 : p. 431-454.

COLLINS, Michael (1996), « A New Statistical Parser Based on Bigram Lexical Dependencies », Proceedings of the 34th ACL, Santa Cruz.

DEKHTYAR, Michael ; DIKOVSKY, Alexander (2004), « Categorial Dependency Grammars », Proceedings of the Conference on Categorial Grammars, Montpellier, France, p. 76-91.

DIKOVSKY, Alexander (2004), « Dependencies as Categories », Proceedings of Recent Advances in Dependency Grammars, COLING'2004 Workshop, p. 90-97.

Dowty, D.R., WALL, R.E, PETERS, S. (1981), *Introduction to Montague semantics*, Linguistics and Philosophy, Reidel.

DUDAU-SOFRONIE, Daniela ; TELLIER, Isabelle (2004a), « A Study of Learnability of Lambek Grammars from Typed Examples », Proceedings of the Conference on Categorial Grammars, Montpellier, p133-147.

DUDAU-SOFRONIE, Daniela ; TELLIER, Isabelle (2004b), « Un modèle d'acquisition de la syntaxe à partir d'informations sémantiques », actes de TALN 04, Fes, p.137-146.

DUDAU-SOFRONIE, Daniela ; TELLIER, Isabelle ; TOMMASI, Marc (2001a), « From Logic to Grammars via Types », Proceedings of the 3rd Workshop on LLL, Strasbourg, p. 35-46.

DUDAU-SOFRONIE, Daniela ; TELLIER, Isabelle ; TOMMASI, Marc (2001b), « Learning Categorial Grammars from Semantic Types », Proceedings of the 13th Amsterdam Colloquium, p. 79-84.

DUDAU-SOFRONIE, Daniela ; TELLIER, Isabelle ; TOMMASI, Marc (2003), « A Learnable Class of CCG from Typed Examples », Proceedings of the 8th Conference on Formal Grammars, Vienna, Austria, p. 77-88.

FORET, Annie ; LE NIR, Yannick (2002), « Rigid lambek grammars are not learnable from string », Proceedings of the 19th International Conference

COLING, Taipei, Taiwan.

GOLD, E. M. (1967), « Language identification in the limit », *Information and Control*, 10 : p. 447-474.

HOCKENMAIER, Julia (2003), *Data and models for statistical parsing with Combinatory Categorial Gramma*, PhD thesis of the University of Edinburgh.

KANAZAWA, Makoto (1998), *Learnable Classes of Categorial Grammars*, Studies in Logic, Language and Information. Stanford, California.

LAMBEK, Joachim (1958), « The Mathematics of Sentence Structure », *American Mathematical Monthly*, 65 : p. 154-170.

LAMBEK, Joachim (1997), « Type grammars revisited », Proceedings of LACL, LNAI 1582, Springer-Verlag, 1999, p. 1-27.

LE NIR, Yannick (2003), *Structures des analyses syntaxiques catégorielles. Application à l'inférence grammaticale*, thèse de doctorat de l'Université de Rennes 1.

MONTAGUE, Richard (1974), *The collected papers of Richard Montague,* Yale University Press.

MOORTGAT, Michael (1997), *Categorial type logic*, Handbook of Logic and Language, p. 93-177. North Holland Elsevier.

MOOT, Richard (2001), « A short introduction to grail », Proceedings of Methods for modalities 2.

MOOT, Richard (2003), « Parsing corpus-induced type logical grammars », Proceedings of the CoLogNet/ElsNet Workshop on Linguistic Corpora and Logic Based Grammars Formalisms.

MOOT, Richard. (2006), « Automated Extraction of Type-Logical Supertags from the Spoken Dutch Corpus », in Bangalore, S. & Joshi, A. (eds), *The Complexity of Lexical Descriptions and its Relevance to Natural Language Processing: A Supertagging Approach*, MIT Press.

MOREAU, Erwan (2004a), « Apprentissage partiel de grammaires lexicalisées », *TAL*, 45-3: p. 71-102.

MOREAU, Erwan (2004b), « Partial learning using link grammars data », Proceedings of ICGI, LNAI 3264, Springer-Verlag, p. 211-222.

MOREAU, Erwan (2005), « Learnable Classes of General Combinatory Grammars », Proceedings of LACL, LNAI 3492, Springer-Verlag, p. 189-

204.

OSBORNE, Miles ; BRISCOE, Ted (1997), « Learning Stochastic Categorial Grammars », Proceedings of the 35th ACL, CoNLL97 Workshop, Madrid, Spain, p. 80-87.

PARTEE, Barbara (1990), *Mathematics methods in linguistics*, Linguistics and Philisophy 30, Kluwer.

POUPARD, Eric ; BÉCHET, Denis ; FORET, Annie (2006), « Acquisition d'une grammaire catégorielle depuis un corpus d'arbre en français », actes de CAP, Presses universitaires de Grenoble, p. 393-394.

PINKER, Steven (1994), *The Language Instinct*, Londres : Penguin Press.

SHINOHARA, Takeshi (1991), « Inductive Inference of Monotonic Formal Systems from Positive Data », *New Generation Computing* 8-4 : p. 371-384.

SLEATOR, Daniel ; TEMPERLEY, Davy (1991), « Parsing English with a Link », Technical report, CMU-CS-91-196, Carnegie Mellon University.

STEEDMAN, Mark (1987), « Combinatory grammars and parasitic gaps », *Natural Language and Linguistic Theory* 5 : p. 403-439.

TELLIER, Isabelle (1999), « Towards a Semantic-based Theory of Language Learning », Proceedings of the 12th Amsterdam Colloquium, p. 217-222.

TIEDE, Hans-Jorg (2001), « Lambek calculus proofs and tree automata », Proceedings of LACL, LNAI 2014, Springer-Verlag.

ZETTLEMOYER, Luke S. ; COLLINS, Michael (2005), « Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorial Grammars », Proceedings of UAI-05.