Apprentissage symbolique de grammaires et traitement automatique des langues

Erwan MOREAU
LINA - FRE 2729, Université de Nantes
2 rue de la Houssinière, BP 92208, F-44322 Nantes cedex 3
Erwan.Moreau@univ-nantes.fr

Résumé. Le modèle de Gold formalise le processus d'apprentissage d'un langage. Nous présentons dans cet article les avantages et inconvénients de ce cadre théorique contraignant, dans la perspective d'applications en TAL. Nous décrivons brièvement les récentes avancées dans ce domaine, qui soulèvent selon nous certaines questions importantes.

Abstract. Gold's model formalizes the learning process of a language. In this paper we present the advantages and drawbacks of this restrictive theoretical framework, in the viewpoint of applications to NLP. We briefly describe recent advances in the domain which, in our opinion, raise some important questions.

Mots-clés : apprentissage symbolique, modèle de Gold, grammaires catégorielles.

Keywords: symbolic learning, Gold's model, categorial grammars.

1 Introduction

L'apprentissage symbolique automatique de grammaires pour les langues naturelles est un domaine relativement méconnu, assez peu étudié et très peu avancé sur le plan applicatif. Cet état de fait s'explique assez facilement : tout d'abord, il s'agit d'une tâche extrêmement complexe, aussi bien dans sa définition précise que dans sa mise en œuvre. Ce sont surtout les aspects théoriques qui en sont étudiés, et il semble jusqu'à présent très difficile d'y obtenir des résultats pratiques dignes d'intérêt (pour le langage naturel). D'un point de vue optimiste, la relative lenteur à passer du stade de l'étude théorique au stade des applications dans ce domaine s'explique par sa grande complexité. En ce sens, ce domaine serait simplement encore trop jeune scientifiquement, mais pourrait prendre de l'ampleur à l'avenir une fois que les bases en seront bien établies. Mais d'un point de vue pessimiste, la complexité excessive de la tâche peut être vue tout simplement comme un obstacle rédhibitoire à d'éventuelles applications.

Pourtant ce domaine est potentiellement riche en applications, si toutefois on admet l'hypothèse quelque peu idéaliste selon laquelle il est possible de construire un algorithme d'apprentissage « parfait ». Celui-ci serait donc capable de donner une grammaire précise d'un langage naturel, pourvu qu'on lui fournisse un nombre suffisant de phrases appartenant à celui-ci. Sous cette hypothèse, la première application (et la plus évidente) est l'analyse syntaxique, elle-même utilisée sous différentes formes dans de nombreux outils de traitement des langues. On pourrait

alors envisager de construire assez facilement des analyseurs, y compris pour des langues pour lesquelles peu d'études linguistiques existent. On peut également penser à coupler l'analyse et l'apprentissage, de façon à mieux prendre en compte la catégorie syntaxique des mots inconnus de l'analyseur. D'autres applications liées à l'analyse, telles que la correction orthographique et syntaxique, sont également susceptibles de bénéficier des apports de l'apprentissage automatique. Si l'on dispose aussi des moyens permettant de gérer l'aspect sémantique des langues, l'autre grande application de l'apprentissage est la génération (passage du sens d'un énoncé à sa réalisation dans une langue précise). Celle-ci est elle-même proche du problème de la traduction automatique, qui est bien sûr une application d'une très grande utilité.

L'*inférence grammaticale* désigne la problématique qui consiste à apprendre des langages à partir de données. Tout cadre formel pour ce problème doit donc avant tout définir les termes *apprentissage*, *langages* et *données*, c'est-à-dire répondre aux questions suivantes : nature des données dont on dispose ? simples séquences de mots (chaînes), arbres, termes, graphes ou tout autre type de structures, mais aussi quantité, qualité, complétude des données. Type de langages considéré, et représentation des langages ? restrictions éventuelles, niveau d'abstraction (e.g. sans contrainte particulière sur la relation entre langages et grammaires, ou au contraire formalisme grammatical précis). Nature du processus d'inférence ? Fini ou non, Solution unique ou multiple, processus automatique ou semi-automatique, résultat précis ou approximation, limites éventuelles sur le temps ou le nombre d'essais.

Le modèle d'identification à la limite, aussi appelé du nom de son auteur modèle de Gold, est l'une des principales représentations formelles du processus d'apprentissage. La première définition en est donnée dans (Gold, 1967). L'auteur lui-même est d'abord pessimiste quant à l'intérêt de ce modèle, à cause de l'apparente impossibilité d'y obtenir des résultats positifs pour des classes de langages « intéressantes ». Plus tard, les résultats positifs obtenus par Angluin dans ce modèle montreront sa pertinence (Angluin, 1980). Le modèle sera ensuite étudié plus en détail : ainsi, plusieurs autres résultats encourageants viendront soutenir l'idée que l'identification à la limite constitue bien un cadre théorique adapté à la représentation du processus d'apprentissage, en particulier celui des langages, voire des langues naturelles.

La question de la pertinence du modèle de Gold par rapport à l'acquisition humaine du langage est plutôt bien étudiée au niveau linguistique et cognitif (Johnson, 2004), mais cette même question est assez peu discutée dans la perspective de l'apprentissage symbolique automatique. C'est pourquoi nous proposons ici une relecture des principaux résultats liés à ce modèle, vu sous l'angle du traitement automatique des langues. Dans cet article nous essaierons donc d'expliquer de façon concise et claire les bases, les outils et les enjeux du modèle de Gold par rapport au TAL. Nous proposons ensuite un point de vue particulier sur les intérêts et limites des résultats obtenus jusqu'ici dans ce cadre, en tentant de donner un peu de recul à cette modeste étude. L'objectif de cet article est donc aussi de soumettre à la discussion quelques questions relatives au domaine, qui nous semblent pertinentes compte tenu de ses récentes évolutions.

2 Identification à la limite

Le principe de l'identification à la limite est la convergence : à partir d'une séquence infinie d'éléments qui caractérisent le langage à deviner, l'apprenant émet des hypothèses. Ces hypothèses prennent la forme d'une grammaire, censée correspondre au langage observé jusqu'alors par l'apprenant. Comme l'énumération est infinie, l'apprenant répond lui aussi sous forme d'une

séquence infinie de grammaires hypothèses. Finalement, l'apprentissage est réussi si, à partir d'un certain point, l'apprenant émet toujours la même hypothèse (convergence), et que celle-ci correspond bien au langage attendu. Le fait que l'apprenant ignorera toujours s'il a atteint ou non la solution est un aspect important de ce formalisme. Gold en donne la justification (d'ordre linguistique) suivante : *« une personne ne sait jamais si elle parle correctement un langage. »*.

Une *classe de langages* est un ensemble de langages[1] fixé, parfois aussi appelé *famille de langages*. Généralement il s'agit d'un ensemble de langages partageant une propriété particulière. L'*apprenabilité*[2] d'une classe de langages désigne son aptitude à être *apprise* selon la définition 2.1 ci-dessous.

Un *système de grammaires* est spécifié par un triplet $\langle \mathcal{U}, \mathcal{G}, \mathcal{M} \rangle$, dans lequel l'univers $\mathcal{U}$ est un ensemble d'objets, $\mathcal{G}$ un ensemble de grammaires et $\mathcal{M}$ une fonction qui associe à chaque grammaire de $\mathcal{G}$ un sous-ensemble de $\mathcal{U}$. Dans un système de grammaires $\langle \mathcal{U}, \mathcal{G}, \mathcal{M} \rangle$, une *fonction d'apprentissage* est une fonction partielle ϕ qui associe à des séquences finies non-vides d'objets de $\mathcal{U}$ des grammaire de $\mathcal{G}$.

Définition 2.1 (Identification à la limite) *Soit $\langle \mathcal{U}, \mathcal{G}, \mathcal{M} \rangle$ un système de grammaires, ϕ une fonction d'apprentissage et $L \subseteq \mathcal{U}$ un langage. Soit $\langle a_0, a_1, a_2, \ldots \rangle$ une séquence infinie d'objets de $\mathcal{U}$, telle que $a \in \{ a_i \mid i \in \mathbb{N} \}$ si et seulement si $a \in L$.*

ϕ converge *vers G s'il existe $n \in \mathbb{N}$ tel que pour tout $i \geq n$ $G_i = \phi(\langle a_1, a_2, \ldots, a_i \rangle)$ est définie et $G_i = G$.*

ϕ apprend *un langage L si, pour toute énumération de L, ϕ converge vers une grammaire G telle que $\mathcal{M}(G) = L$.*

Une classe de langages $\mathcal{L} \subseteq \mathcal{P}(\mathcal{U})$ est dite apprenable *s'il existe une fonction d'apprentissage ϕ telle que ϕ apprend L pour tout langage $L \in \mathcal{L}$.*

On voit dans cette définition que la séquence d'exemples a quelques caractéristiques notables :

- Elle ne contient que des exemples *positifs*, c'est-à-dire des éléments du langage. La fonction d'apprentissage n'a donc aucune information extérieure au langage, ce qui constitue la principale difficulté de cette forme d'apprentissage.
- La séquence d'exemples est supposée ne comporter aucune erreur[3].
- La séquence d'exemples est une *énumération* du langage : tous les objets du langage doivent obligatoirement y apparaître.
- Les exemples peuvent apparaître dans un ordre quelconque dans la séquence, et éventuellement plusieurs fois (ce qui permet notamment d'énumérer indéfiniment un langage fini).

Dans ce modèle, la convergence d'une fonction d'apprentissage n'a d'intérêt que si elle s'applique à un ensemble de langages, et non à un seul langage. Intuitivement, plus la classe de langages est grande, plus il est difficile de reconnaître un langage précis dans cette classe.

[1] Dans la littérature, le terme *langage* est fréquemment défini comme un ensemble de phrases, chaque phrase étant une séquence finie de mots. Mais dans la mesure où on peut envisager différents niveaux de représentation de la phrase, nous définissons un *langage* comme un ensemble d'*objets* (ce terme abstrait laissant volontairement la possibilité d'utiliser différents types d'éléments : arbres, structures, etc.).

[2] On trouve aussi dans la littérature différents termes désignant le caractère apprenable d'une classe de langages : *inférable*, *identifiable [à la limite]* ou *acquiérable* (le terme *acquisition* fait cependant plus souvent référence à l'apprentissage humain du langage).

[3] Ce qui limite les applications potentielles : ce modèle est par définition inadapté aux données bruitées.

Cette définition de l'identification à la limite a d'importantes conséquences immédiates, démontrées par Gold dans (Gold, 1967). La première est un résultat positif pour les langages finis : La classe des langages de cardinalité finie est apprenable. En effet, intuitivement il suffit dans ce cas que l'apprenant ajoute un par un les exemples présentés à la grammaire hypothèse : la fonction d'apprentissage converge dès que le langage a été énuméré en totalité. En revanche la seconde conséquence de la définition du modèle est un résultat négatif : toute classe de langages contenant tous les langages finis et au moins un langage infini n'est pas apprenable. Nous illustrons ce résultat à l'aide de l'exemple ci-dessous :

Exemple 2.1 (Langages réguliers) *Pour tout $n \geq 1$ on définit $L_n = \{ x^i \mid i \leq n \}$ comme le langage des chaînes de x de longueur inférieure ou égale à n. Soit $L_\infty = x^*$ le langage de toutes les chaînes de x.*

Supposons que la séquence d'exemples commence par $\langle x, xx, xxx, \ldots \rangle$:
- *Si l'apprenant est prudent, il ne propose jamais comme hypothèse un langage qui va au delà des exemples proposés : il propose donc L_k, avec k la longueur maximale parmi les exemples vus. Cet algorithme ne peut jamais trouver L^∞.*
- *Si à l'inverse l'algorithme « généralise », alors à partir d'un certain point il propose L^∞. Mais c'est une erreur s'il s'avère que la séquence ne dépasse pas une certaine longueur de phrase.*

Une erreur est donc possible dans les deux cas, et rien ne permet de faire le bon choix : si la classe de langage contient tous les L^n et L^∞, celle-ci n'est pas apprenable.

Les langages $\{ L_n \mid n \in \mathbb{N} \}$ et L_∞ définis dans l'exemple 2.1 étant tous réguliers, toutes les classes de la hiérarchie de Chomsky les contiennent, et ne sont par conséquent pas apprenables. Le fait que même la classe la plus simple de la hiérarchie de Chomsky, celle des langages réguliers, ne soit pas apprenable dans le modèle de Gold a longtemps constitué un obstacle majeur au développement du modèle, considéré comme trop contraignant. Gold lui-même notait : *« Cependant, les résultats présentés dans la dernière section montrent que seule la classe de langages la plus triviale* [4] *considérée*[5] *est apprenable à partir d'exemples positifs[..]. »*

L'exemple 2.1 illustre la principale difficulté de l'apprentissage à partir d'exemples positifs, à savoir la *surgénéralisation*. La surgénéralisation est l'erreur qui consiste à trop généraliser (extrapoler) à partir des données fournies, ce qui signifie inférer un langage qui est un sur-ensemble strict du langage cible. Par exemple, on peut supposer que l'ensemble $\{11, 23, 5, 17, 7\}$ est le début d'une énumération de l'ensemble des nombres impairs. Mais s'il s'agit en fait de l'ensemble des nombres premiers supérieurs à 2, alors il y a surgénéralisation : l'ensemble des nombres représenté est un sous-ensemble (strict) de l'ensemble proposé. Comme on ne dispose que d'exemples positifs, il n'y aura jamais de contre-exemple dans la séquence permettant de corriger l'erreur. Bien entendu, la généralisation est indispensable dans le processus d'apprentissage, puisqu'une méthode d'apprentissage « trop prudente » qui ne généraliserait jamais ne ferait pas de véritable *apprentissage* (au sens d'une découverte de quelque chose de nouveau) : il s'agirait simplement d'une sorte de compilation des exemples proposés. Surtout, il est évident qu'une telle méthode serait incapable d'identifier un langage infini.

Sauf cas particuliers, la généralisation doit donc bien être utilisée au cours de l'apprentissage. La question qui se pose d'un point de vue algorithmique est : quand faut-il généraliser ? (ou

[4] On peut considérer de manière informelle qu'une classe de langages est *non triviale* (pour l'apprentissage) si elle comporte au moins un nombre infinis de langages, dont certains sont infinis.

[5] Il s'agit de la classe des langages de cardinalité finie.

quand faut-il ne pas généraliser, selon ce qu'on considère comme étant l'action par défaut). Mais avant de se poser cette question, il faut s'assurer qu'il est *possible de savoir quand généraliser*, car lorsqu'on ne dispose pas d'exemples négatifs on n'a aucun indice sur la position de la frontière du langage à deviner. C'est précisément ce point qui pose problème au départ avec l'identification à la limite à partir d'exemples positifs : même les classes de langages qu'on croyait simples (les langages réguliers, voir exemple 2.1) ne sont pas apprenables, parce qu'on ne peut pas savoir quand [ne pas] généraliser.

3 Techniques théoriques d'apprentissage de grammaires

3.1 Ensembles révélateurs

Au début des années 1980, Angluin apporte au modèle de Gold ses premiers résultats positifs : dans (Angluin, 1980), elle propose de *« considérer le cas particulier d'inférence à partir d'exemples positifs qui évite la surgénéralisation [et donne] des conditions suffisantes pour cela. »* Le critère qu'elle propose a donné lieu ensuite à de nombreuses utilisations ou extensions, démontrant finalement la richesse du modèle de Gold. Sommairement, un ensemble révélateur *(telltale set)* est une sorte de « signature » d'un langage qui le distingue de tous les autres langages de la classe dont il est un sur-ensemble strict. Ainsi, lorsque cette signature apparaît dans la séquence d'exemples, on peut proposer ce langage sans risque de surgénéralisation. Formellement, soit $\mathcal{L}$ une classe de langages : un ensemble fini d'objets D est un *ensemble révélateur* du langage $L \in \mathcal{L}$ si $D \subseteq L$ et $L' \subset L \Rightarrow D \nsubseteq L'$ pour tout langage $L' \in \mathcal{L}$.

Théorème 3.1 (Angluin) *Soit $\mathcal{L} \subseteq \mathcal{P}(\mathcal{U})$ une famille indexée de langages récursifs*[6] *dans le système de grammaires $\langle \mathcal{U}, \mathcal{G}, \mathcal{M} \rangle$: $\mathcal{L} = \{ \mathcal{M}(G) \mid G \in \{G_0, G_1, G_2, \ldots\} \}$.*

Il existe une fonction ϕ qui apprend $\mathcal{L}$ si et seulement s'il existe un algorithme calculable qui, pour tout indice I tel que $L_I = \mathcal{M}(G_I) \in \mathcal{L}$ énumère un ensemble révélateur de L_I.

Supposons qu'il existe un algorithme $EnumRevel(I, n)$ qui énumère récursivement les n premiers éléments d'un ensemble révélateur D_I de L_I.

$\phi(\langle a_0, \ldots, a_n \rangle)$
  $i \leftarrow 0$
  $E \leftarrow EnumRevel(i, n)$
  Tant que ($i \leq n$ **et non**($\{a_0, \ldots, a_n\} \subseteq \mathcal{M}(G_i)$ **et** $E \subseteq \{a_0, \ldots, a_n\}$)) **faire**
    $i \leftarrow i + 1$
    $E \leftarrow EnumRevel(i, n)$
  Fin Tant Que
  Renvoyer G_i

L'algorithme ci-dessus apprend la classe $\mathcal{L}$ de la façon suivante : un ensemble révélateur étant nécessairement fini, pour tout i il existe une étape n à partir de laquelle l'ensemble révélateur E de L_i est énuméré en totalité. Dans ce cas, la boucle s'arrêtera sur la première grammaire G_i telle que les exemples fournis appartiennent au langage de la grammaire d'une part, et dont

[6] Classe de langage pour laquelle le problème de l'appartenance ($x \in L$) est décidable.

l'ensemble révélateur est entièrement inclus dans les exemples d'autre part. Ainsi il est impossible que le langage cible soit un sous-ensemble strict de $\mathcal{M}(G_i)$ (surgénéralisation). Si aucun contre-exemple $a_j \notin \mathcal{M}(G_i)$ n'apparaît par la suite, l'algorithme s'arrêtera toujours sur G_i.

L'apprentissage par énumération, illustré ci-dessus, désigne une méthode générale qui consiste à faire une recherche systématique dans l'ensemble des grammaires possibles, jusqu'à en trouver une qui vérifie une propriété particulière, et la renvoyer comme hypothèse. L'énumération n'est pas une méthode d'apprentissage universelle, parce qu'il existe des classes de langages apprenables qui ne sont pas apprenables par énumération (Costa Florêncio, 2003). Il va de soi que ce type d'algorithme est totalement inutilisable en pratique : même si on peut améliorer sensiblement la méthode (notamment en évitant de faire l'énumération complète après chaque exemple), le seul fait d'avoir à parcourir de façon exhaustive l'ensemble des grammaires potentielles est rédhibitoire. En effet, imaginons une représentation textuelle simple des grammaires (de type grammaires syntagmatiques), de façon à les énumérer selon l'ordre de taille puis lexicographique : en première approximation il existe de l'ordre de n^m grammaires différentes de taille m, avec n le nombre total de symboles (comprenant entre autres tous les mots du vocabulaire). Ce type de fonctionnement est évidemment radicalement inadapté à l'apprentissage de langues naturelles[7]. Il faut donc souligner cette caractéristique essentielle du modèle de Gold : il s'agit avant tout d'un modèle *théorique*, qui ne garantit que la *décidabilité* du problème de l'apprentissage. Autrement dit, le fait qu'une classe de langages soit apprenable n'implique en aucun cas la faisabilité pratique (en temps raisonnable) du processus d'apprentissage.

3.2 Élasticité finie

L'élasticité [in]finie est une propriété définie par (Motoki *et al.*, 1991) de la façon suivante : Soit $\langle \mathcal{U}, \mathcal{G}, \mathcal{M} \rangle$ un système de grammaires. Une classe $\mathcal{L} \subseteq \mathcal{P}(\mathcal{U})$ a l'élasticité infinie s'il existe une séquence infinie $\langle a_0, a_1, a_2, \ldots \rangle$ d'objets dans $\mathcal{U}$ et une séquence infinie $\langle L_1, L_2, \ldots \rangle$ de langages dans $\mathcal{L}$ tels que $a_i \notin L_i$ et $\{a_0, \ldots, a_{i-1}\} \subseteq L_i$ pour tout $i > 0$. Une classe de langages $\mathcal{L}$ a la propriété d'*élasticité finie* si elle n'a pas l'élasticité infinie. L'élasticité finie est donc une condition suffisante pour l'apprenabilité. De fait, il s'agit d'une propriété très utile pour démontrer l'apprenabilité de nouvelles classes de langages, car cette condition est souvent plus simple à vérifier que l'existence globale d'un algorithme convergent. L'élasticité finie est notamment utilisée par Shinohara pour définir une nouvelle condition suffisante à l'aide du concept de *densité finie bornée*, qui lui permet de démontrer l'apprenabilité de la classe des grammaires syntagmatiques contextuelles d'au plus k règles (pour tout $k \geq 0$) (Shinohara, 1991). De plus, Kanazawa a démontré une propriété très pratique, qui permet de montrer l'élasticité finie (donc aussi l'apprenabilité) d'une classe de langages complexe à partir du cas d'une classe plus simple possédant la propriété (voir ci-dessous). D'un point de vue algorithmique, on notera que tous les résultats d'apprenabilité obtenus à l'aide de l'élasticité finie reposent finalement sur l'algorithme d'apprentissage par énumération des ensembles révélateurs (présenté plus haut).

Théorème 3.2 (Kanazawa) *Soient $\mathcal{U}$ et $\mathcal{U}'$ deux ensembles d'objets, et $\mathcal{L}$ une classe de langages définie sur $\mathcal{U}$ qui a l'élasticité finie. S'il existe une relation $R \subseteq \mathcal{U}' \times \mathcal{U}$ finiment valuée, alors la classe de langages $\mathcal{L}' = \{ R^{-1}[L] \mid L \in \mathcal{L} \}$ a aussi l'élasticité finie*[8].

[7] Notons que l'acquisition humaine du langage n'a certainement rien à voir non plus avec cette méthode.

[8] Une relation binaire R sur $A \times B$ est *finiment valuée* si et seulement si pour tout $a \in A$ il n'existe qu'un nombre fini de $b \in B$ tels que $a \, R \, b$. Si L est un langage sur $\mathcal{U}$ et R une relation sur $\mathcal{U}' \times \mathcal{U}$, l'*image inverse* de L par rapport à R est le langage $R^{-1}[L] = \{ a \in \mathcal{U}' \mid \exists b \in L \text{ tel que } a \, R \, b \}$.

4 Apprentissage de grammaires catégorielles

En 1998, Kanazawa propose plusieurs résultats importants concernant l'apprenabilité des grammaires AB dans le modèle de Gold (Kanazawa, 1998). Les grammaires AB, la forme la plus simple de grammaires catégorielles, sont (totalement) lexicalisées : à chaque mot sont associés un ou plusieurs types syntaxiques dans le lexique (règles *lexicales*), et deux règles *universelles* définissent la façon dont ces types peuvent se combiner entre eux dans les dérivations[9].

Les apports de Kanazawa sont multiples : il montre de nouveaux résultats et développe de nouvelles techniques de preuve. Surtout, ses résultats sont les premiers pour le modèle de Gold à traiter d'un formalisme grammatical présentant certaines prédispositions à la représentation des langues naturelles, à savoir les grammaires catégorielles. Plus précisément, les grammaires AB sont assez pauvres sur le plan de la représentation linguistique. Mais la famille des grammaires catégorielles contient d'autres formalismes beaucoup plus puissants pour représenter des langues naturelles, c'est pourquoi le premier résultat prometteur de Kanazawa a donné lieu à d'autres travaux visant à étendre l'apprenabilité à des formes plus riches de grammaires.

4.1 Apprenabilité des grammaires AB

Parmi ses résultats, il faut distinguer deux aspects très différents du point de vue applicatif :

Il y a tout d'abord un aspect algorithmique, basé sur l'algorithme d'apprentissage RG proposé dans (Buszkowski & Penn, 1989). Cet algorithme apprend efficacement des grammaires AB rigides[10] à partir de FA-structures. Ces structures sont une forme « d'arbre de dérivation appauvri » des phrases, c'est-à-dire qu'elles ne contiennent pas toutes les informations d'un arbre de dérivation (sans quoi il n'y aurait aucun apprentissage, puisque les types seraient déjà donnés), mais tout de même beaucoup plus d'information que de simples chaînes : parenthésage des constituants, ainsi qu'une forme particulière d'orientation des dépendances entre constituants. Il est donc plus facile d'apprendre lorsqu'on dispose en plus de cette information structurée.

Le second aspect concerne l'apprenabilité d'une classe de langages plus étendue. Kanazawa ne montre pas seulement l'apprenabilité de la classe des langages de FA-structures de grammaires AB rigides, il démontre aussi que cette classe a l'élasticité finie. Or grâce au théorème 3.2, il prouve que cette propriété est également vérifiée par la classe des langages de chaînes des grammaires AB k-valuées[11] (pour tout $k \geq 0$), donc cette classe est elle aussi apprenable. Ce résultat est beaucoup plus intéressant pour deux raisons : d'une part la contrainte de rigidité est levée, ce qui permet d'envisager de représenter un langage naturel avec ces grammaires[12]. D'autre part il n'est plus nécessaire de disposer des FA-structures avec les exemples de phrases, ce qui est un avantage important puisque celles-ci constituent une information spécifique au formalisme, en pratique très difficile à obtenir en quantité. En revanche, on ne dispose pas dans ce cas d'algorithme d'apprentissage efficace[13].

[9] Voir par exemple (Moreau, 2006) pour une définition complète.

[10] Une grammaire est rigide si à chaque mot du vocabulaire n'est associé qu'un seul type syntaxique.

[11] Une grammaire est k-valuée si à chaque mot du vocabulaire n'est associé qu'au plus k types différents.

[12] La rigidité empêche en effet toute forme d'homonymie. Mais surtout elle ne permet pas de représenter de manière satisfaisante la plupart des mots grammaticaux, car leur usage syntaxique prend souvent des formes variés.

[13] Au contraire, Costa-Florêncio démontre qu'il s'agit d'un problème NP-dur (Costa Florêncio, 2003).

4.2 Extensions à d'autres formalismes

Les bons résultats obtenus par Kanazawa avec les grammaires AB posent la question de savoir si les grammaires catégorielles ont certaines propriétés qui feraient d'elles de bonnes candidates à l'apprentissage dans le modèle de Gold. Cette question du formalisme grammatical est importante pour d'éventuelles applications aux langues naturelles, puisque celles-ci nécessitent une représentation à la fois linguistiquement fiable et aussi utilisable le plus facilement possible. En ce qui concerne l'apprenabilité efficace à partir de structures (de type FA-structures, mais la forme peut varier selon les formalismes), plusieurs résultats viendront montrer ensuite que ce type d'apprentissage peut être étendu à d'autre formalismes sans grande difficulté. Kanazawa donne lui-même l'exemple des grammaires combinatoires générales (GCG). Des résultats équivalents sont obtenus avec différents formalismes, notamment les grammaires de Lambek et les grammaires minimalistes (Bonato & Retoré, 2001), mais toujours au prix d'une contrainte similaire à la rigidité (limitations sur le nombre ou la forme des règles lexicales associées à un mot), et toujours avec l'aide d'informations structurelles assez précises.

Mais le passage du cas « grammaires rigides et avec structures » au cas « grammaires k-valuées ou sans structure », qui constitue le point fort des résultats de Kanazawa, s'avère nettement plus difficile lorsqu'on s'éloigne du cas des grammaires AB. On aurait pu supposer que les propriétés logiques des grammaires AB jouaient un rôle pour l'apprenabilité, mais cette hypothèse est invalidée par les résultats négatifs des grammaires de Lambek (Foret & Le Nir, 2002). Une autre hypothèse de travail a consisté à considérer les grammaires AB comme un système de grammaires (lexicalisées) spécifié par un ensemble particulier de règles universelles (de réécriture par substitution). On peut alors étudier ce qui les distingue des autres systèmes dans le cadre plus large des GCG proposé par Kanazawa (Moreau, 2006) : on cherche ainsi des conditions, portant sur les règles universelles, qui sont suffisantes pour l'apprenabilité des classes de langages correspondantes (on espère trouver de cette manière des ensembles de règles plus fines qui permettent l'apprenabilité). En se basant sur la méthode employée par Kanazawa, nous avons ainsi montré que certaines classes de GCG ont l'élasticité finie (donc sont apprenables) : les *grammaires à arguments bornés* k-valuées, qui représentent des classes de langages assez vastes, mais souffrent d'une limitation « technique » (sur la taille des arguments) difficile à justifier au niveau linguistique. Les *grammaires par consommation stricte d'arguments* k-valuées sont en revanche apprenables sans limitation, mais sont définies par un critère tellement strict qu'on ne s'éloigne pas beaucoup du cas des grammaires AB. De plus, il ne s'agit pas que d'une limite « temporaire » (c'est-à-dire susceptible d'être repoussée à l'avenir) car les *grammaires par consommation d'arguments* (rigides), qui en sont un sur-ensemble très peu élargi, n'ont pas l'élasticité finie : cela signifie qu'on atteint ici, entre ces deux cas relativement proches, les frontières de l'apprenabilité des GCG (du moins selon la méthode de Kanazawa).

4.3 Applications à l'apprentissage symbolique du langage naturel ?

Compte tenu des contraintes du modèle et des résultats présentés ci-dessus, il est compréhensible que les applications de ce type d'apprentissage au langage naturel demeurent très modestes. De fait, le premier problème à résoudre est cette équation apparemment insoluble : soit on cherche à apprendre à l'aide d'informations structurées, mais le type d'information requis n'existe pas en quantité suffisante a priori ; ou bien on tente d'apprendre à partir de simples phrases, mais alors on ne dispose que d'algorithmes de complexité exponentielle, incapables de réaliser le processus en temps raisonnable.

Différentes méthodes ont été envisagées, qui font toutes appel à des ressources structurées, de façon plus ou moins directe. Quelques unes utilisent des corpus de structures spécifiques, obtenus manuellement ou par conversion plus ou moins automatique de ressources arborées existantes (Dudau-Sofronie, 2004). Nous avions également proposé une approche intermédiaire, à partir de chaînes mais avec l'apport d'un sous-ensemble de la grammaire cible (Moreau, 2006), en utilisant un lexique existant sous forme de grammaires de liens. Le fait qu'il soit nécessaire de faire appel à des ressources externes, souvent exprimées dans un formalisme grammatical particulier, pose un problème théorique de fond du point de vue du modèle de Gold : où s'arrête la notion d'inférence grammaticale, c'est-à-dire d'apprentissage symbolique de la syntaxe, et où commence la « simple » extraction d'informations syntaxiques ? En effet, l'usage de ressources externes facilite bien sûr l'apprentissage, mais introduit aussi un biais dans le processus : à partir d'un certain niveau d'informations syntaxiques fournies, il ne s'agit plus d'apprentissage mais de reconstitution de la grammaire qui a servi à produire les exemples, qu'elle soit formellement établie ou sous-jacente. On risque alors de ne faire que reproduire des schémas syntaxiques préétablis, la grammaire résultante n'aurait donc pas beaucoup d'intérêt : dans ce cas elle peut être construite directement de façon semi-automatique, à partir des règles qui ont défini la création des données. Un autre travers plus subtil peut également apparaître : le simple étiquetage syntaxique par des catégories prédéfinies (nom, verbe, adjectif, etc.) est une forme appauvrie d'apprentissage de la syntaxe, car ce cadre empêche de tenir compte d'éventuelles variations par rapport aux catégories de départ. Dans ce cas, il n'est pas certain que le modèle de Gold ait quelque chose de plus à apporter au problème que les techniques existantes en TALN.

5 Conclusion

L'acquisition automatique de grammaires ne se limite pas à l'apprentissage (au sens de Gold). Par exemple, pour certaines formes de grammaires catégorielles, les travaux d'Hockenmaier (Hockenmaier, 2003) ou de Moot (Moortgat & Moot, 2001) montrent qu'il est possible d'obtenir une grammaire à large couverture d'un langage naturel, à partir de corpus structurés. Mais leur approche est à notre sens plus proche de l'*extraction* automatique que de l'inférence grammaticale, car dans les deux cas des techniques ad hoc de conversion des données sont utilisées.

L'utilisation d'un modèle contraignant comme le modèle de Gold constitue une garantie de « précision » de la grammaire obtenue, parce qu'il donne une direction générale au processus de l'apprentissage : l'existence d'un objectif (qu'on peut considérer comme idéal) définissant *ce que doit être* la grammaire apprise diffère de la simple extraction d'information syntaxique, dans laquelle on obtient toujours un résultat (quelles que soient les données), et ce résultat n'est justifié qu'a posteriori (parfois selon une évaluation spécifique, souvent simplement par son utilité). Typiquement, le problème de la surgénéralisation est difficile voire impossible à détecter dans le cas de l'extraction, tandis que le modèle de Gold impose d'en tenir compte *a priori* dans l'algorithme d'apprentissage (sans quoi la convergence ne serait pas vérifiée).

Il est vrai que le modèle de Gold est avant tout un modèle théorique, et le critère de convergence sur lequel il repose ne semble pas vraiment approprié pour des applications de traitement automatique. Dans (Angluin & Smith, 1983), Angluin concluait son état de l'art sur l'inférence inductive par la remarque suivante : *« Le problème ouvert le plus important n'est sans doute pas une quelconque question technique spécifique, mais le fossé entre les résultats abstraits et concrets. »* Force est de constater que, malgré quelques progrès indéniables sur le plan théorique, les tentatives d'applications concrètes de cette forme d'apprentissage restent encore peu

concluantes, parce qu'on ne parvient pas à (on ne peut pas ?) apprendre sur des données réelles sans relâcher tout ou partie des contraintes du modèle. Cela ne signifie pas nécessairement que l'on perde ainsi tout l'intérêt du modèle, mais dans ces conditions il nous semble judicieux de redéfinir l'objectif de la tâche d'apprentissage : inférence grammaticale, extraction, ou approche mixte ? Étant donné les difficultés rencontrées lorsqu'on s'en tient strictement au modèle, cette dernière possibilité semble la plus réaliste.

Toutefois, peut-être que l'algorithme d'apprentissage idéal n'est tout simplement pas encore découvert : dans ce cas, *« les générations futures riront bien de notre ignorance actuelle. »* (Angluin & Smith, 1983).

Références

ANGLUIN D. (1980). Inductive inference of formal languages from positive data. *Information and Control*, **48**, 117–135.

ANGLUIN D. & SMITH C. H. (1983). Inductive inference : Theory and methods. *ACM Computing Surveys*, **15**(3), 237–269.

BONATO R. & RETORÉ C. (2001). Learning rigid lambek grammars and minimalist grammars from structured sentences. In *Proc. of 3d Workshop on Learning Language in Logic*, p. 23–34.

BUSZKOWSKI W. & PENN G. (1989). *Categorial grammars determined from linguistic data by unification*. Rapport interne TR-89-05, Dpt of Computer Science, University of Chicago.

COSTA FLORÊNCIO C. (2003). *Learning categorial grammars*. PhD thesis, Utrecht University.

DUDAU-SOFRONIE D. (2004). *Apprentissage de grammaires catégorielles pour simuler l'acquisition du langage naturel à l'aide d'informations sémantiques*. PhD thesis, Univ. Lille 1.

FORET A. & LE NIR Y. (2002). Lambek rigid grammars are not learnable from strings. In *COLING'2002, 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

GOLD E. (1967). Language identification in the limit. *Information and control*, **10**(5), 447–474.

HOCKENMAIER J. (2003). *Data and models for statistical parsing with Combinatory Categorial Grammar*. PhD thesis, School of Informatics, The University of Edinburgh.

JOHNSON K. (2004). Gold's theorem and cognitive science. *Philosophy of Science*, **71**, 571–592.

KANAZAWA M. (1998). *Learnable classes of categorial grammars*. Cambridge University Press.

MOORTGAT M. & MOOT R. (2001). CGN to Grail : Extracting a type-logical lexicon from the CGN annotation. In *Proceedings of CLIN 2000* : W. Daelemans.

MOREAU E. (2006). *Acquisition de grammaires lexicalisées pour les langues naturelles*. PhD thesis, Université de Nantes.

MOTOKI T., SHINOHARA T. & WRIGHT K. (1991). The correct definition of finite elasticity : corrigendum to Identification of unions. In *Proceedings of the Fourth Annual Workshop on Computational Learning Theory*, p. 375, San Mateo, CA : Morgan Kaufmann.

SHINOHARA T. (1991). Inductive inference of monotonic formal systems from positive data. *New Generation Computing*, **8**(4), 371–384.