Modélisation des processus d'acquisition syntaxique par jeux de langage entre agents artificiels

Marie Marcia[1] Isabelle Tellier[1]
(1) Lattice (UMR 8094), CNRS, ENS Paris, Université Sorbonne Nouvelle,
PSL Research University, USPC, 1 rue Maurice Arnoux, 92120 Montrouge, France
marie.marcia@univ-paris3.fr, isabelle.tellier@univ-paris3.fr

RÉSUMÉ

Dans cet article, nous présentons une modélisation de la situation d'acquisition de la syntaxe de sa langue maternelle par un enfant inspirée des "jeux de langages" de Luc Steels. Le modèle suppose que l'enfant a accès à une représentation sémantique des énoncés qui lui sont adressés, et qu'il doit réagir en désignant la tête syntaxique de ces énoncés. Nous décrivons des expériences exploitant des données du corpus CHILDES et mettant en jeu un processus d'acquisition simple mais efficace.

ABSTRACT

Modeling Syntactic Acquisition by Language Games between Artificial Agents

In this paper, we present a model of the way a child acquires her mother tongue syntax, inspired by Luc Steels' "language games". The model assumes that the child has access to the semantics of the utterances addressed to her, and that she must react by designing the syntactic head of these utterances. We describe experiments made with data from the CHILDES database, using a simple but effective acquisition process.

MOTS-CLÉS : analyse syntaxique, sémantique, modélisation de l'acquisition, jeux de langage.

KEYWORDS: parsing, semantics, model of language acquisition, language games.

1 Introduction

Nous proposons dans cet article un modèle d'apprentissage automatique de la syntaxe d'une langue naturelle qui s'inspire de l'acquisition humaine. Ce modèle s'appuie sur le paradigme des jeux de langage, qui visent à modéliser l'émergence de connaissances linguistiques dans une population de locuteurs artificiels. Le cadre proposé ici fait intervenir non pas une population mais seulement deux locuteurs artificiels : un expert (ou adulte) et un apprenant (ou enfant) qui doit acquérir, au fil d'interactions avec l'expert, une compétence syntaxique. Les composantes phonétique et lexicale de l'apprentissage de la langue sont considérées comme déjà acquises [1] et n'entreront pas en ligne de compte. La sémantique joue en revanche un rôle fondamental. Notre modèle s'appuie sur des capacités sémantiques conçues comme un prérequis nécessaire à l'acquisition syntaxique.

La section suivante définit le paradigme des jeux de langage dans lequel se place notre modèle. Nous

1. Il ne s'agit pas de suggérer qu'à l'âge où l'enfant commence à acquérir des compétences syntaxiques, les compétences phonétique et lexicale sont complètement acquises. Cependant, la tâche proposée portera uniquement sur l'acquisition de la syntaxe, où la connaissance lexicale notamment est un prérequis.

détaillerons ensuite le protocole expérimental mis en place : le corpus utilisé, le déroulement du jeu de langage défini, les méthodes de parsing et d'apprentissage automatique mis en œuvre. Enfin, les résultats des premières expériences menées seront analysés et évalués.

2 Jeux de langage

Une approche originale de l'acquisition de connaissances linguistiques a été proposée par le paradigme des jeux de langage (Steels, 1995; Kaplan, 2001). Ce paradigme, au sein de la linguistique dite évolutionnaire (Kirby, 2002; Dessalles, 2000), tente de rendre compte des processus pouvant mener à l'émergence et à l'évolution des langues naturelles, en proposant des expérimentations logicielles et robotiques. Puisqu'il est difficile d'observer empiriquement l'émergence d'une nouvelle langue, les chercheurs utilisant les jeux de langage suggèrent de recréer artificiellement les conditions d'émergence d'un système linguistique : une communauté de locuteurs potentiels, un objectif communicatif et des sens à exprimer. À partir de ces éléments, ils proposent des mécanismes pouvant mener à l'émergence d'un lexique commun.

L'intérêt majeur du paradigme des jeux de langage est de trouver un équilibre entre la simplicité de la modélisation et la complexité de l'interaction langagière située (Wellens, 2012). Ce paradigme repose sur la simulation d'interactions langagières simplifiées entre binômes de locuteurs artificiels (robotiques ou logiciels), qui tentent de s'exprimer (et de se comprendre) à propos d'un contexte d'objets (réels ou également artificiels). On peut comparer une simulation de ce type à l'émergence d'un protolangage (Bickerton, 1990) dans une population humaine, qui correspond à un stade pré-syntaxique de la langue où s'établit une correspondance entre des unités lexicales et des unités sémantiques. Ce type d'émergence et d'évolution au sein d'un groupe est à mettre en parallèle avec l'acquisition de la langue par l'enfant, qui connaît également ce stade pré-syntaxique.

Le paradigme des jeux de langage s'est surtout focalisé jusqu'à présent sur l'acquisition de connaissances lexicales. Plus récemment (Steels & Garcia-Casademont, 2015), il a été étendu à des jeux de langage syntaxiques. Il s'agit alors, à partir d'une simulation, de faire émerger une grammaire commune chez une population d'agents disposant déjà de capacités lexicales et d'une représentation sémantique du monde. Cette grammaire émerge, là encore, grâce à une succession d'interactions des locuteurs s'exprimant à propos de leur monde. La grammaire se construit peu à peu par la réalisation de jeux de langage consistant en la production et l'interprétation d'énoncés référant à des objets ou des événements du monde simulé.

On trouve dans ce paradigme l'essentiel des paramètres que nous voulons réaliser dans notre propre modélisation : un modèle évolutif basé sur l'interaction, l'utilisation d'exemples positifs de la langue à acquérir et la composante sémantique nécessaire à l'acquisition d'une grammaire. Mais notre modèle se distingue de celui de Steels par le formalisme syntaxico-sémantique utilisé et le nombre d'agents intervenant dans l'interaction : une population chez Steels, une simple paire "expert/apprenant" pour nous. Notre modélisation cherche en effet à simuler l'acquisition individuelle de la langue par un enfant et non l'émergence d'une langue nouvelle dans un groupe.

3 Protocole expérimental

3.1 Corpus

Le corpus que nous avons constitué pour l'input fourni par l'expert à l'apprenant est extrait d'une collection des corpus français de la base de données de la plate-forme CHILDES (MacWhinney, 2000). Ces textes sont la transcription de discussions entre des enfants (cibles d'études parfois longitudinales) et d'autres interlocuteurs (les parents, les linguistes ou des pairs). Ces données sont particulièrement adaptées à notre tâche puisqu'elles constituent un corpus de Français oral [2] qui comporte des milliers de phrases adressées (pour la majorité) par des adultes à des enfants.

Nous avons extrait du corpus tous les énoncés produits par les participants autres que l'enfant cible, puis filtré ces énoncés pour qu'il ne reste plus que ceux syntaxiquement analysables. Le parser utilisé pour cela est Grew (Guillaume & Perrier, 2012), qui présente l'intérêt de produire aussi une représentation sémantique des énoncés (cf. section 3.2). Le critère minimaliste utilisé est que le parser renvoie un seul graphe de dépendances, donc que la phrase ne comporte qu'une seule racine. Les autres phrases, non analysables par le parser, sont éliminées. Avec ce tri, environ 44% des phrases sont éliminées. Le parser Grew est en effet conçu pour traiter du Français écrit et non oral. La proportion de phrases analysables est cependant amplement suffisante pour notre corpus.

La Table 1 donne le nombre de phrases prononcées par des adultes extraites du corpus en fonction de l'âge de l'enfant destinataire. Les proportions étant très disparates, le corpus a été échantillonné pour obtenir une égale proportion d'inputs présentés à des enfants de 0 à 1 an (11 mois et 17 jours pour le plus jeune), de 1 à 2 ans, etc, jusqu'à 6-7 ans (6 ans 11 mois et 26 jours pour le plus âgé) : 400 énoncés par tranche d'âge.

Âge de l'enfant cible	0	1	2	3	4	5	6
Nombre de phrases	745	55363	102688	44244	15062	6283	2621
Longueur moyenne des phrases	4.57	4.88	5.61	5.68	6.04	5.93	5.7

TABLE 1 – Nombre de phrases et longueur moyenne des phrases en fonction de l'âge de l'enfant destinataire dans le corpus non échantillonné

Le tableau donne aussi la longueur moyenne des phrases adressées aux enfants, qui augmente en fonction de leur tranche d'âge [3]. Les 2800 énoncés constituant le corpus échantillonné ont donc été triés pour apparaître, pendant nos expériences, par ordre croissant d'âge de l'enfant destinataire.

Nous donnons pour information la représentation des différentes catégories morphosyntaxiques (selon le jeu d'étiquettes morphosyntaxiques de MElt (Denis & Sagot, 2009) utilisé ici) parmi les racines syntaxiques de la totalité des énoncés de notre corpus. Il s'agit pour la majorité de racines verbales, mais on trouve également une proportion considérable de noms, prépositions et conjonctions (de coordination ou de subordination) :

— Verbes : 64.11%

2. Voici quelques exemples de phrases que l'on pourra trouver dans notre corpus : "encore une banane", "ah bah le hérisson", "ah hop", "tu me disais", "non non non", "pas prendre le cube d'Ulysse", "ohlàlà Julie".

3. La longueur des énoncés est certes à mettre en lien avec l'âge de l'enfant destinataire, mais la façon de parler des individus est évidemment variable et influe fortement sur cette longueur moyenne. Le type de dialogue peut également avoir une influence : le discours des parents à leur enfant est plutôt spontané, tandis que certaines questions posées à l'enfant par le chercheur peuvent être préparées, et éventuellement plus longues.

— Noms : 12.11%
— Prépositions : 9.82%
— Conjonctions : 8.82%
— Mots étrangers [4] : 2.07%
— Adverbes : 1.07%
— Pronoms : 0.93%
— Adjectifs : 0.57%
— Prépositions + déterminants : 0.50%

3.2 Déroulement du jeu de langage

Dans notre modèle interviennent seulement deux interlocuteurs artificiels, un expert et un apprenant (ou bien un adulte et un enfant). L'apprenant n'a initialement aucune connaissance syntaxique, mais il a des connaissances lexicales. Il est donc capable de comprendre le sens de mots isolés. Considérer la connaissance lexicale comme acquise est un parti pris qui se base premièrement sur le fait que, chronologiquement, le processus d'acquisition d'une compétence lexicale chez l'enfant intervient avant le début du processus d'acquisition syntaxique. Il existe un chevauchement dans l'acquisition de ces deux compétences, mais la compétence syntaxique permettant de comprendre et de former des "phrases" de plus d'un mot apparaît bel et bien après l'étape de l'acquisition lexicale qui permet à l'enfant d'associer un sens à un nombre de mots encore limité (c'est-à-dire vers l'âge de 2 ans). On donne dans le tableau suivant, tiré de (Tellier, 2005) la chronologie de l'acquisition des différentes compétences linguistiques. D'autre part, notre travail se focalise sur l'acquisition de la syntaxe : l'apprentissage du lien entre un mot et son sens est une tâche distincte que nous préférons écarter en le considérant comme acquis.

Au cours d'un jeu, donc d'une interaction, l'expert "prononce" un énoncé (provenant du corpus échantillonné). L'apprenant a accès à cet énoncé et à sa représentation sémantique globale (sous forme de graphe [5]). On considère en effet qu'un énoncé est toujours prononcé dans un contexte dans lequel il peut être compris. Le graphe sémantique de l'énoncé simule ce contexte virtuel, réduit au minimum puisque limité à cet énoncé. L'apprenant est donc exposé à un énoncé au sein duquel il est capable de faire le lien entre un mot et son sens lexical, ainsi qu'au sens global de cet énoncé. Mais il n'a pas accès à sa structure syntaxique.

Le but du jeu est, pour l'apprenant, de désigner, dans son environnement (le contexte, représenté par le graphe sémantique), le sens (donc le nœud sémantique) exprimé par la racine syntaxique de l'énoncé. Ce principe s'inspire d'environnements simulés (par exemple SHRDLU (Winograd, 1971)) où les objets sont désignés par un terme qui est la racine d'une phrase nominale ("le triangle bleu sur le cube rouge"). Nous étendons ce principe à des énoncés oraux où la racine peut désigner autre chose qu'un objet. La notion de désignation devient alors plus conceptuelle. Dans une phrase comme "Elle est là la maison du singe", l'apprenant doit "désigner" le sens correspondant au verbe.

Pour réussir le jeu, l'apprenant doit acquérir une compétence syntaxique. Une fois que l'enfant a donné sa réponse, l'expert confirme ou infirme mais ne donne pas la solution. Le but n'est pas en effet que l'enfant apprenne à associer une bonne réponse à une phrase donnée, mais qu'il apprenne à

4. Cette catégorie recouvre en fait à la fois les mots étrangers, mais aussi les onomatopées et marqueurs de l'oral comme "tchou", "bah", "ben", "okay", "hou", etc.

5. Le graphe sémantique est un graphe orienté, connecté, acyclique, et n'a pas de racine. On peut voir en Figure 3 qu'un nœud sémantique peut aussi bien n'avoir aucun gouverneur qu'en avoir un ou plusieurs.

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

FIGURE 1 – Chronologie de l'acquisition des compétences linguistiques

induire une structure syntaxique à partir de n'importe quelle phrase. L'expérience consiste en une série d'interactions de ce type, qui testent et simulent uniquement la capacité de compréhension.

La réussite d'un jeu de désignation en compréhension est soumise à la nécessité de tenter une analyse de la structure syntaxique de l'énoncé. On pourrait créer un jeu de désignation en production, où les rôles seraient inversés : l'apprenant devrait produire des énoncés et l'enseignant en désigner la racine syntaxique. Si l'enseignant parvient à trouver la bonne racine, c'est que l'énoncé est suffisamment grammatical. L'intérêt en termes d'apprentissage serait moindre puisqu'il s'agirait pour l'apprenant de produire une séquence de mots, et non une structure syntaxique. Le feedback de l'enseignant apporterait également peu à l'apprenant : soit l'enseignant confirme ou infirme la grammaticalité de la séquence (ce qui n'apporte pas de connaissance syntaxique à l'apprenant), soit il fournit une correction (or ce n'est pas l'objectif de notre modélisation de donner une correction ou "feedback négatif"). De manière générale, notre dispositif se prête mal à des jeux de production par l'apprenant : si le jeu réussit, il n'apprenant rien, s'il échoue, il obtient une éventuelle correction.

D'autres types de jeux de compréhension sont cependant envisagés pour la suite de ce travail : le premier est un jeu, non plus de désignation, mais dans lequel l'apprenant doit déterminer la valeur de vérité d'un énoncé produit par l'adulte ; le second introduit des phrases interrogatives produites par l'enseignant et auxquelles l'apprenant doit répondre par oui ou par non en fonction du contexte.

La connaissance de l'expert (l'analyse syntaxique juste) et le graphe sémantique d'une phrase sont obtenus grâce à un outil de parsing par réécritures de graphes, Grew. Celui-ci permet, à partir d'une phrase étiquetée en POS (ici par MElt (Denis & Sagot, 2009)), d'obtenir un graphe de dépendances de cette phrase, ainsi que sa représentation sémantique en DMRS (Dependency Minimal Recursion Semantics) (Copestake, 2009), représentation dont l'exigence est "d'avoir une annotation lisible et

minimale" (Guillaume & Perrier, 2012).

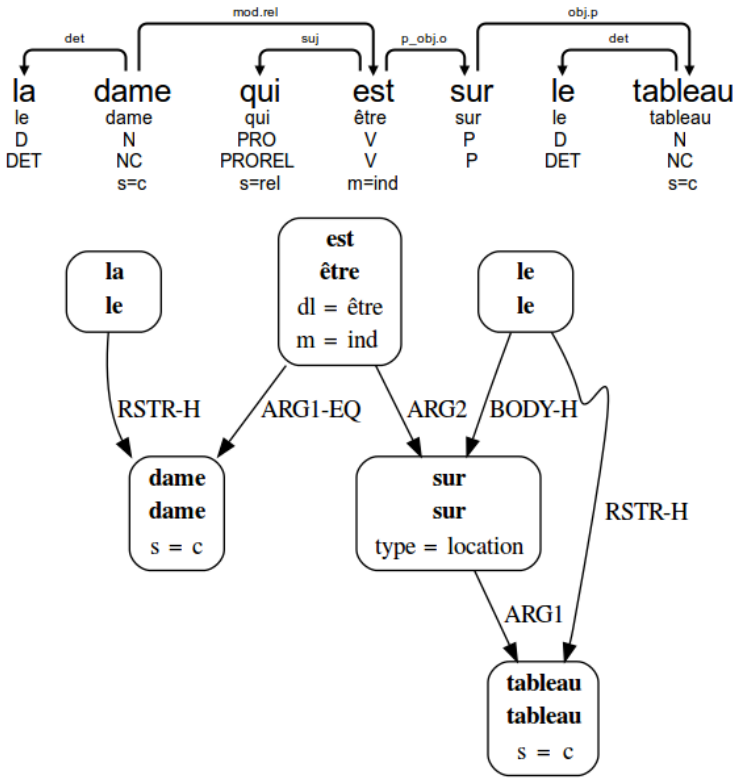

FIGURE 2 – Analyse en dépendances (en haut) et représentation sémantique (en bas) d'un énoncé du corpus réalisées grâce au parser Grew

Les Figures 2 et 3 donnent des exemples d'énoncés du corpus avec leur analyse en dépendances et leur graphe sémantique associés. Dans le cas de la Figure 2, l'expert prononce l'énoncé "la dame qui est sur le tableau". L'apprenant, après analyse de l'énoncé, doit désigner le sens correspondant à la racine syntaxique de l'énoncé. La racine étant le token "dame", l'apprenant doit désigner le noeud sémantique correspondant à ce token. On simule par ce procédé un jeu de désignation d'objet où l'enfant doit comprendre de quel objet on parle, et le désigner. Dans une expérience comme SHRDLU, l'apprenant aurait ici la possibilité de désigner la dame ou le tableau. Dans notre expérience, il a la possibilité de désigner n'importe lequel des sens exprimés dans le graphe sémantique ("la", "est", "le", "dame", "sur" ou "tableau"). Dans la Figure 3, on applique le même principe, qui devient alors plus conceptuel, puisque l'apprenant doit à nouveau désigner le sens correspondant à la racine syntaxique de l'énoncé, le token "est", or ce token ne désigne plus un objet. Si la tâche de désignation devient plus conceptuelle, elle permet en revanche de ne pas avoir à trier le corpus pour n'en conserver que des phrases nominales (qui correspondent mieux à une tâche de désignation classique). Ainsi, l'apprenant est exposé à un input plus varié et plus réaliste dans une tâche d'apprentissage de la syntaxe. Des énoncés comme "ohlàlà Julie" ou "encore une banane" (présents dans notre corpus) se prêtent moins à une tâche de désignation, mais sont bien des énoncés auxquels des enfants ont été exposés, et qui ont participé à leur acquisition de la syntaxe du Français. Le biais du jeu de désignation est un moyen de donner un objectif à l'interaction, objectif qui ne peut être atteint que par l'intermédiaire de l'apprentissage de la syntaxe.

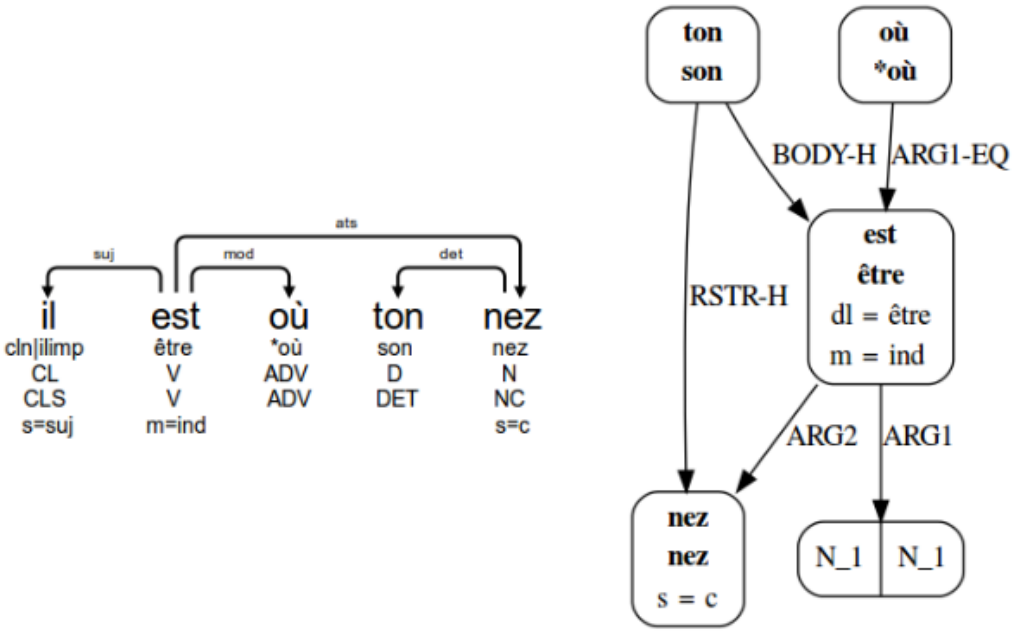

FIGURE 3 – Analyse en dépendances (à gauche) et représentation sémantique (à droite) d'un énoncé du corpus réalisées grâce au parser Grew

3.3 Parsing

Dans notre modélisation, l'analyse d'un énoncé par l'apprenant est produite par un parser en dépendances de type shift-reduce inspiré de (Yamada & Matsumoto, 2003). On modélise donc l'acquisition par l'apprentissage automatique non pas d'une grammaire (comme dans (Tellier, 2005)) mais d'un parser. Le parsing par shift-reduce revient à opérer une succession de classifications, tout en gérant une pile. C'est un modèle simple qui peut être acquis incrémentalement, et donc une hypothèse cognitivement raisonnable pour l'acquisition. Le classifieur traite des paires de tokens de la phrase et leur associe une classe. Cette classe correspond à une décision syntaxique : soit on ne met aucun lien de dépendance entre les deux tokens et on passe à la paire suivante (classe No Link), soit on met un lien de dépendance vers la gauche et on supprime le token en haut de la pile (classe Left Arc), soit on crée un lien vers la droite et on passe à la paire suivante (classe Right Arc). Le type de shift-reduce choisi ici permet d'établir des liens de dépendance non étiquetés, qui ne présupposent donc aucune connaissance syntaxique de la part de l'apprenant.

Quand, au cours du shift-reduce, toutes les paires disponibles dans la pile ont déjà été analysées, on stoppe le shift-reduce, et on analyse à la suite toutes les autres paires de tokens possibles de la phrase qui n'ont pas encore été analysées, et dont les deux éléments sont encore présents dans la pile (dans le shift-reduce, si un élément a été "réduit", donc supprimé de la pile, c'est qu'il a déjà un gouverneur et qu'il ne peut pas avoir de dépendants).

Pour classer les paires de tokens (n1, n2), l'apprenant analyse la séquence de l'énoncé et sa représentation sémantique, et construit un vecteur. Les traits choisis pour la représentation vectorielle des paires de tokens sont : l'écart entre les positions des deux tokens dans la phrase, leurs lemmes, ceux des contextes gauche et droit de n1 et n2 ; de quels arguments les nœuds sémantiques correspondant aux tokens n1 et n2 sont les prédicats, et quels sont les types de ces liens ; de quels prédicats ils sont les arguments, et le type de ces liens ; si les nœuds correspondant à n1 et n2 sont sémantiquement liés,

par un lien vers la gauche ou vers la droite. L'utilisation de l'étiquetage en POS comme trait de la représentation vectorielle, prépondérante dans les tâches d'apprentissage d'un analyseur syntaxique, est exclue ici puisque l'apprenant ne dispose d'aucune information syntaxique a priori.

Lorsqu'à un token ne correspond pas de nœud sémantique (ce qui concerne certains mots grammaticaux), les traits du vecteur correspondant aux informations sémantiques liées à ce token ont des valeurs nulles. L'absence d'information sémantique dans la représentation vectorielle de ce type de token étant commune aux mots grammaticaux, on obtient donc une classe de vecteurs particuliers (représentant une paire de tokens dont l'un ou les deux sont des mots grammaticaux) qui auront entre eux une distance vectorielle réduite. En effet, l'apprentissage des mots grammaticaux se distingue de celui des mots lexicaux : en l'absence de sens associé aux mots grammaticaux, c'est plutôt leur contexte d'apparition (dans la séquence de l'énoncé) qui va servir à leur intégration dans l'acquisition des structures syntaxiques.

Une fois la paire analysée et représentée sous la forme d'un vecteur, il faut ensuite choisir sa classe. Cette décision est prise par classification automatique avec l'algorithme des k plus proches voisins (avec une valeur de k à optimiser). Cette technique présente plusieurs avantages : elle est incrémentale et ne nécessite que peu de calcul. Très simple, elle pose peu d'hypothèses sur l'acquisition. Elle permet en outre d'utiliser des données à la fois symboliques et numériques, et ne nécessite pas de connaître à l'avance le vocabulaire présent dans l'input. Une fois que la décision par knn est prise, on vérifie que le lien à créer ne va pas donner un second gouverneur à un token.

Une fois la phrase entièrement analysée par le parser courant de l'apprenant, on choisit une racine au hasard (s'il y en a plusieurs) parmi celles produites. On la compare alors avec celle de l'enseignant (analysée avec le parser de Grew) qui donne un feedback (bonne ou mauvaise réponse). Enfin, on met à jour les données d'entraînement. On traite d'abord les nouvelles données issues de l'énoncé en entrée. Une donnée est un vecteur qui contient des informations sur une paire de tokens (avec les informations sémantiques associées) : ce vecteur se voit attribuer un score de confiance pour chaque classe. Si la donnée a déjà été rencontrée, on met à jour les scores des classes par inhibition latérale (voir paragraphe suivant). Si la donnée n'a jamais été rencontrée, on initialise les scores des classes en fonction du feedback (si le feedback est positif, les décisions qui ont été prises lors du parsing ont un meilleur score que les décisions qui n'ont pas été retenues, et inversement). On met ensuite à jour les scores des vecteurs proches des entrées rencontrées, par inhibition latérale.

La notion d'inhibition latérale est issue de la neurobiologie (Hartline *et al.*, 1956) et réfère à la capacité d'un neurone actif à réduire l'activité de ses voisins (Wellens, 2012)). Elle a été introduite dans les jeux de langage par (Steels & Kaplan, 1999). Dans un jeu de langage comme le "Naming Game" (Steels, 2000), au cours duquel un groupe d'agents artificiels doit inventer des mots pour référer à des objets puis converger vers un lexique commun pour désigner ces objets au fil d'interactions par binômes d'agents, la stratégie de l'inhibition latérale est efficace pour résoudre le problème de la concurrence entre plusieurs mots pour désigner un même objet. Plus un agent entend, au cours de ses interactions successives, un mot *m* pour désigner un objet, plus le score exprimant la certitude que le mot *m* désigne bien cet objet va augmenter, tandis que le score des mots concurrents pour désigner cet objet va décroître. Ici, la mise à jour des scores par inhibition latérale consiste, dans le cas de la réussite du jeu, à augmenter le score de la classe choisie et diminuer les scores des autres classes (et inversement dans le cas d'un échec du jeu), en fonction du feedback de l'enseignant. Autrement dit, si le jeu réussit, l'apprenant renforce sa certitude quant aux décisions syntaxiques qui ont été prises au cours de l'interaction, tandis que décroît sa certitude quant aux décisions qu'il n'a pas prises (et vice-versa en cas d'échec du jeu). Les formules de mise à jour des scores provenant de (Wellens,

2012) font intervenir les paramètres de renforcement r et d'inhibition i, tous deux fixés à 0.3 dans nos expériences (mais optimisables), et le score initial s d'une classe pour un vecteur. On utilise cette formule pour renforcer un score : $s + r \,.\, (1 - s)$; et la suivante pour inhiber un score : $s - i \,.\, s$. Les socres sont toujours compris entre 0 et 1.

4 Résultats et analyses

4.1 Mesure de réussite communicative

La mesure de réussite communicative est propre aux jeux de langage. Elle se calcule en divisant le nombre de fois où le jeu est un succès par le nombre total de jeux effectués (on obtient donc un score compris entre 0 et 1). Cependant, la réussite communicative gage que l'apprenant a trouvé la racine de la phrase, mais pas que le graphe de dépendance qu'il a construit pour y parvenir est correct.

Pour nos expériences, chaque séquence de jeux de langage est effectuée dix fois. La valeur optimale trouvée pour k est 5. On donne en Figure 4 les courbes moyenne, maximale et minimale de l'évolution de la réussite communicative. Deux séries d'expériences sont ici présentées : dans l'une, le corpus d'entrées est trié par âge de l'enfant destinataire ; dans l'autre, il est trié par longueur des phrases. Le tri des phrases par âge de l'enfant destinataire permet une modélisation plus réaliste. Le discours des parents étant dans une certaine mesure adapté à l'âge de l'enfant destinataire, les phrases ont donc une complexité croissante, mais cette difficulté est relative plus au lexique qu'à la syntaxe. D'un point de vue formel, le lexique n'étant pas l'objet de l'apprentissage, cette évolution de la difficulté n'a donc pas d'impact sur l'apprentissage du classifieur. Trier le corpus par longueur de phrases, de manière certes artificielle du point de vue de la modélisation, permet néanmoins de donner un critère formel qui introduit une difficulté croissante dans l'apprentissage du classifieur, et améliore ses résultats.

Lorsque le corpus est trié par âge de l'enfant destinataire, la courbe de réussite croît jusqu'à se stabiliser au-dessus de 0.6 en moyenne. Les résultats obtenus sont meilleurs lorsque les phrases de l'expert sont triées par longueur. On observe une augmentation du taux de réussite jusqu'à un palier (0.85 environ), malgré une difficulté croissante. En fin d'expérience, la difficulté des jeux provoque une légère baisse du taux de réussite. Le critère de longueur des phrases est plus formel et engendre une difficulté croissante dans l'expérience qui favorise l'apprentissage.

Nous donnons pour comparaison en Figure 5 les résultats obtenus pour la mesure de réussite communicative dans une baseline où les informations sémantiques relatives aux paires de tokens ne sont pas prises en compte dans leur représentation vectorielle. Lorsque le corpus est trié par âge de l'enfant destinataire, la réussite stagne rapidement autour de 0.45. Les résultats sont encore une fois meilleurs lorsque le corpus est trié par longueur des énoncés. Mais lorsqu'on ne tient pas compte des informations sémantiques, il faut plus de temps à l'apprenant pour atteindre son palier maximal, celui-ci est moins élevé, et l'augmentation de la longueur des énoncés fait décroître la réussite de manière significative. L'information sémantique sémantique facilite donc l'apprentissage et rend le système plus robuste à l'augmentation de la taille des énoncés.

Pour comparaison également est donnée en Figure 6 l'évolution de la réussite communicative pour une baseline où la classification est faite au hasard.

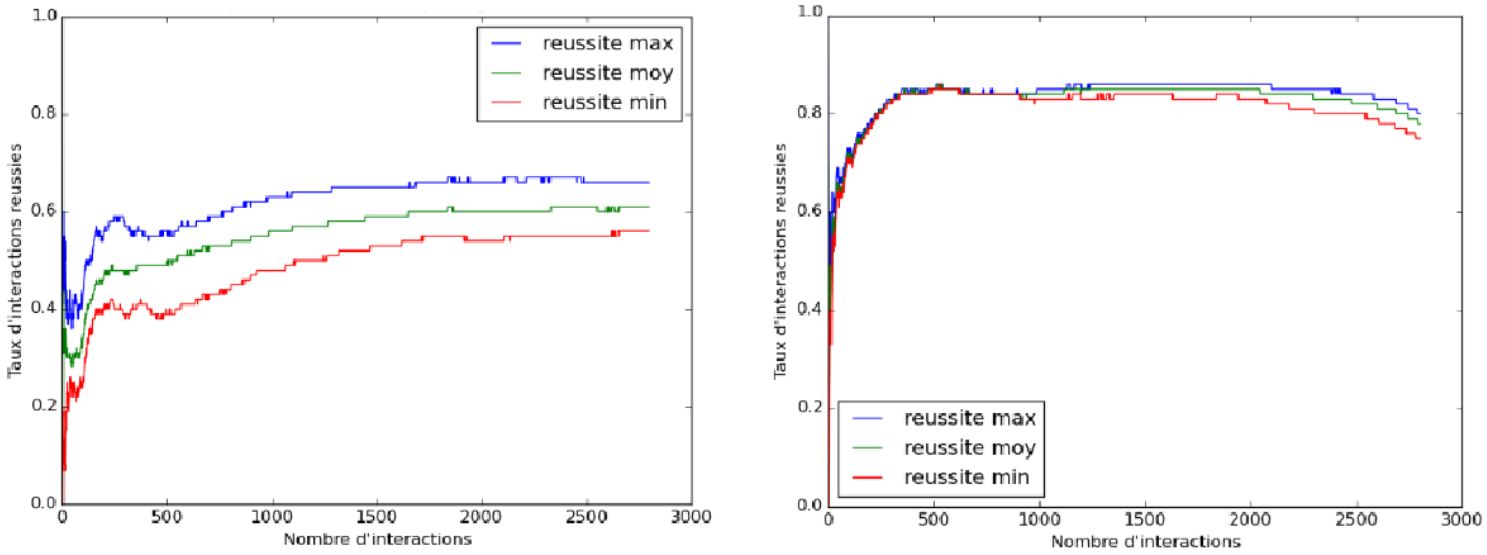

FIGURE 4 – Évolution de la proportion de jeux réussis en fonction du nombre d'interactions. À gauche, les énoncés sont triés par ordre croissant d'âge de l'enfant cible ; à droite, par longueur

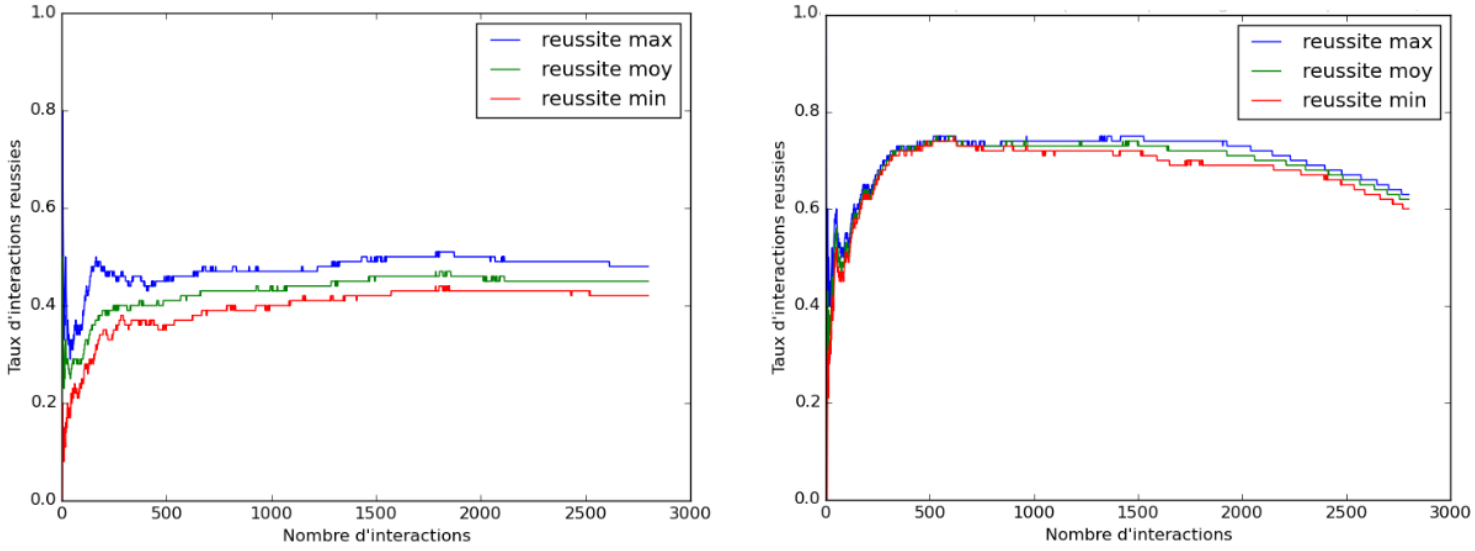

FIGURE 5 – Évolution de la proportion de jeux réussis en fonction du nombre d'interactions, avec une baseline sans informations sémantiques. À gauche, les énoncés sont triés par ordre croissant d'âge de l'enfant cible ; à droite, par longueur

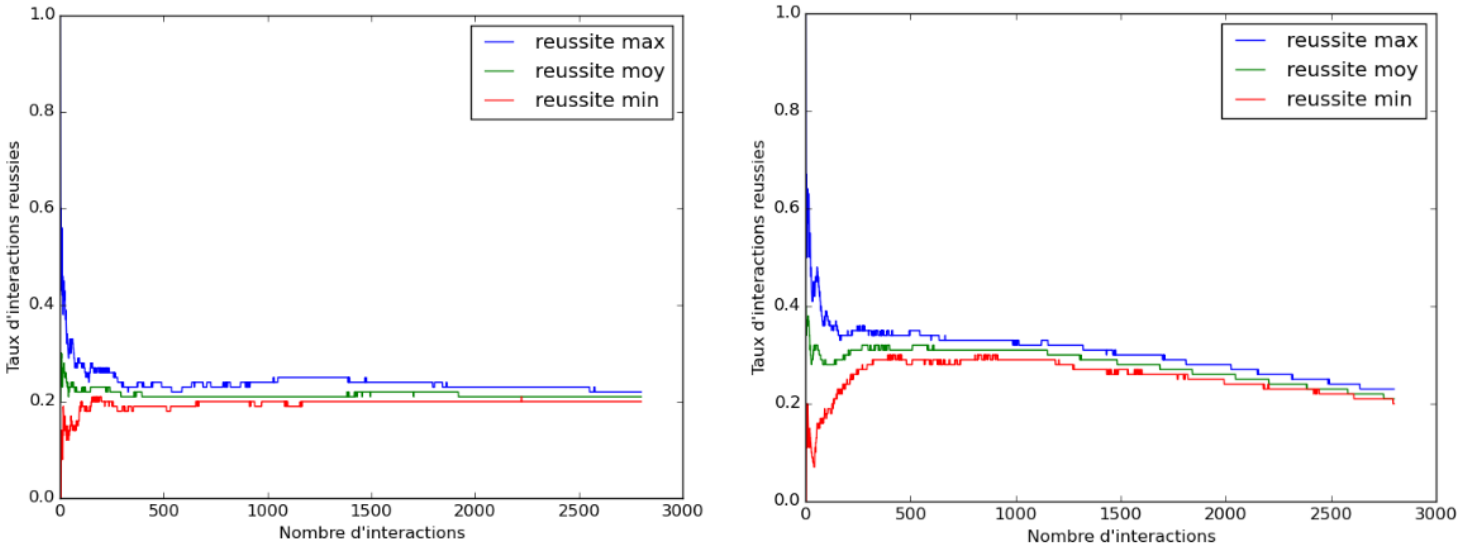

FIGURE 6 – Évolution de la proportion de jeux réussis en fonction du nombre d'interactions, avec une baseline où la classification est faite au hasard. À gauche, les énoncés sont triés par ordre croissant d'âge de l'enfant cible ; à droite, par longueur

4.2 Mesure UAS

La mesure d'UAS (Unlabeled Attachment Score) est plus classique pour évaluer la qualité d'un analyseur syntaxique relativement à un corpus de référence. Elle mesure la proportion de tokens pour lequel le bon gouverneur a été trouvé. Dans notre cas, nous ne disposons pas d'analyses de référence pour nos données, simplement du résultat fourni par l'application de MElt et Grew, qui peuvent faire des erreurs. Comme ce résultat sert de base à la sémantique, qui est utilisée par l'apprenant, c'est néanmoins par rapport à lui que nous nous évaluerons.

Le tableau 2 donne l'évolution de la mesure d'UAS au fil de l'expérience sur le corpus trié par longueur de phrases, avec (expérience standard) et sans utilisation de l'information sémantique (baseline 1), et lorsque la classification est faite au hasard (baseline 2), pour des portions de 400 phrases. On donne également la longueur moyenne des phrases pour chaque portion de corpus, et le nombre moyen de dépendances correctes trouvées (lors de l'expérience standard prenant en compte l'information sémantique).

Phrases	1-400	401-800	801-1200	1201-1600	1601-2000	2001-2400	2401-2800
UAS	0.87	0.78	0.67	0.69	0.61	0.58	0.46
UAS BL1	0.77	0.7	0.57	0.67	0.56	0.55	0.37
UAS BL2	0.33	0.31	0.3	0.29	0.28	0.26	0.26
Long. moy.	2.0	2.46	3.16	4.0	4.88	6.16	8.9
Nb. dép.	1.74	1.92	2.12	2.76	2.98	3.57	4.09

TABLE 2 – Mesure d'UAS (pour l'expérience standard, pour la baseline sans sémantique (BL1) et pour la baseline faisant une classification au hasard (BL2)), longueur moyenne des phrases et nombre moyen de dépendances correctes trouvées par phrase, par portion de 400 phrases du corpus au fil de l'expérience

Plus les phrases sont longues, plus l'UAS diminue (sauf entre les portions 801-1200 et 1201-1600). Malgré une difficulté croissante pour trouver le bon arbre de dépendances, l'apprenant réussit le jeu dans la proportion vue précédemment. On note également que, malgré une baisse de l'UAS, le nombre de dépendances correctes trouvées par l'apprenant augmente tout au long de l'expérience.

5 Conclusion et perspectives

Nous avons présenté un modèle d'apprentissage automatique de la syntaxe qui s'inspire de l'acquisition d'une langue naturelle par les enfants, et qui permet l'apprentissage d'un parser sans corpus arboré, mais avec seulement une représentation sémantique des phrases. Les résultats de ce modèle sont satisfaisants, mais peuvent être améliorés notamment par l'utilisation d'autres types de classifieurs (il est prévu d'implémenter un modèle utilisant des réseaux de neurones). Il doit pouvoir s'adapter à d'autres langues, ce qu'il faudra également tester.

Références

BASSANO D., LAAHA S., MAILLOCHON I. & DRESSLER W. U. (2004). Early acquisition of verb grammar and lexical development : Evidence from periphrastic constructions in french and austrian german. *First Language, 24(1)*, p. 33–70.

BICKERTON D. (1990). *Language and Species*. Chicago : The University of Chicago Press.

CHAMPAUD C. (1994). The development of verb forms in french children at around two years of age : some comparisons with romance and non-romance languages. *First Lisbon Meeting on Child Language*.

COPESTAKE A. (2009). Invited talk : Slacker semantics : Why superficiality, dependency and avoidance of commitment can be the right way to go. p. 1–9.

DEMUTH K. & TREMBLAY A. (2008). Prosodically-conditioned variability in children's production of french determiners. *Journal of Child Language, 35*, p. 99–127.

DENIS P. & SAGOT B. (2009). Coupling an annotated corpus and a morphosyntactic lexicon for state-of-the-art pos tagging with less human effort.

DESSALLES J.-L. (2000). *Aux origines du langage. Une histoire naturelle de la parole*. Paris : Hermes Science.

GOAD H. & BUCKLEY M. (2006). Prosodic structure in child french : Evidence for the foot. *Catalan Journal of Linguistics, 5*, p. 109–142.

GUILLAUME B. & PERRIER G. (2012). Annotation sémantique du french treebank à l'aide de la réécriture modulaire de graphes. *Actes TALN 2012*, p. 293–306.

HAMANN C., OHAYON S., DUBÉ S., FRAUENFELDER U., RIZZI L., STARKE M. & ZESIGER P. (2003). Aspects of grammatical development in young french children with sli. *Developmental Science, 6*, p. 151–159.

HARTLINE K., WAGNER H. & RATLIFF F. (1956). Inhibition in the eye of limulus. *Journal of General Physiology, 39*, p. 651–673.

HUNKELER H. (2005). Aspects of the evolution of the early lexicon in the interactions mother-child : Case study of two dizygotic twin children between 15 and 26 months.

KAPLAN F. (2001). *La naissance d'une langue chez les robots*.

KERN S., DAVIS B. L. & ZINK I. (2009). From babbling to first words in four languages : Common trends, cross language and individual differences. *First Lisbon Meeting on Child Language*.

KIRBY S. M. (2002). Natural language from artificial life. *Artificial Life, 8*, p. 185–215.

LEROY M., MATHIOT E. & MORGENSTERN A. (2009). Pointing gestures and demonstrative words : Deixis between the ages of one and three. p. 386–404.

MACWHINNEY B. (2000). *The CHILDES Project : Tools for analyzing talk. Third Edition*. Mahwah, NJ : Lawrence Erlbaum.

NORMAND M. T. L., MORENO-TORRES I., PARISSE C. & DELLATOLAS G. (2013). How do children acquire early grammar and build multiword utterances ? a corpus study of french children aged 2 to 4. *Child Development, 84(2)*, p. 647–661.

PALASIS K. (2010). *Syntaxe générative et acquisition : le sujet dans le développement du système linguistique du jeune enfant*.

PLUNKETT B. (2003). Null subjects and the setting of subject agreement parameters in child french. *Romance Linguistics : Theory and Acquisition*, p. 351–366.

P.SUPPES, SMITH R. & LEVEILLÉ M. (1973). The french syntax of a child's noun phrases. *Archives de Psychologie, 42*, p. 207–269.

STEELS L. (1995). A self-organizing spatial vocabulary. *Artificial Life*, p. 319–332.

STEELS L. (2000). Language as a complex adaptive system. p. 17–28.

STEELS L. & GARCIA-CASADEMONT E. (2015). How to play the syntax game. *Proceedings of the European Conference on Artificial Life 2015*, p. 479–486.

STEELS L. & KAPLAN F. (1999). Collective learning and semiotic dynamics. *Advances in Artificial Life. Proceedings of the Fifth European Conference, ECAL'99*, p. 679–688.

TELLIER I. (2005). Modéliser l'acquisition de la syntaxe du langage via l'hypothèse de la primauté du sens. HDR d'informatique, université Lille3.

VIHMAN M. M., DEPAOLIS R. A. & DAVIS B. L. (1998). Is there a "trochaic bias" in early word learning ? evidence from english and french. *Child Development, 69*, p. 933–947.

WELLENS P. (2012). *Adaptive Strategies in the Emergence of Lexical Systems*. Vrije Universiteit Brussel, Bruxelles : VUBPRESS Brussels University Press.

WINOGRAD T. (1971). *Procedures as a Representation for Data in a Computer Program for Understanding Natural Language*. Cambrige : Massachusetts Institute of Technology Project MAC Report.

YAMADA H. & MATSUMOTO Y. (2003). Statistical dependency analysis with support vector machines. *Proceedings of IWPT, vol. 3*, p. 195–206.

YAMAGUCHI N. (2012). *Parcours d'acquisition des sons du langage chez deux enfants francophones*. Université Sorbonne Nouvelle Paris 3.