Création automatique d'une grammaire syntaxico-sémantique

Émilie Colin[1]
(1) équipe SYNALP, Université de Lorraine/LORIA, Nancy, France
emilie.colin@loria.fr

Résumé

Nous proposons une nouvelle méthode pour la création automatique de grammaires lexicalisées syntaxico-sémantiques. A l'heure actuelle, la création de grammaire résulte soit d'un travail manuel soit d'un traitement automatisé de corpus arboré. Notre proposition est d'extraire à partir de données VerbNet une grammaire noyau (formes canoniques des verbes et des groupes nominaux) de l'anglais intégrant une sémantique VerbNet. Notre objectif est de profiter des larges ressources existantes pour produire un système de génération de texte symbolique de qualité en domaine restreint.

Abstract

Syntactic-semantic grammar automatic creation

We propose a new method to automaticaly create a syntactico-semantic lexicalized grammar. For the moment, this work is done either manually or with the support of an automated processing of annotated corpus. Our proposal is to extract verbal, nominal and auxiliary trees from VerbNet. Our goal is to take advantage of the vast resources available to produce a system of quality text symbolic generation in a restricted domain.

Mots-clés : grammaire, génération, extraction, ressources.

Keywords: grammar, generation, extraction, resources.

1 Introduction

Une grammaire syntaxico-sémantique décrit la relation entre langue naturelle, syntaxe et sémantique. Couplée avec un algorithme de réalisation de surface, elle peut notamment être utilisée pour générer des phrases (Narayan & Gardent, 2012; Cahill & Van Genabith, 2006; Carroll & Oepen, 2005). Pour une représentation sémantique donnée, l'algorithme de réalisation de surface produira la ou les phrases associée(s) par la grammaire à cette représentation.

Les méthodes existantes pour la création de telles grammaires sont souvent manuelles (Gardent, 2006) ou spécifiques à un domaine donné (Wong & Mooney, 2007; Gyawali, 2016). Dans cet article, nous proposons une méthode qui permet de créer automatiquement une grammaire noyau de l'anglais avec une sémantique générique intégrant les rôles thématiques de Verbnet (Kipper *et al.*, 2008). La grammaire est une grammaire d'arbres adjoints lexicalisée à traits (FB-LTAG, (Vijay-Shanker & Joshi, 1988)) et est créée à partir des exemples contenus

dans VerbNet et de leur annotation syntaxico-sémantique. Afin de faciliter la maintenance et l'extension de cette grammaire (e.g., extension à des formes non canoniques comme les propositions relatives), l'extraction automatique produit une méta-grammaire (Candito, 1998, 1996) plutôt qu'une grammaire. Les fragments d'arbres et de sémantiques produits par l'extraction sont ensuite assemblés par un compilateur de métagrammaires existant (XMG, (Crabbé *et al.*, 2013), section 4) pour créer les arbres de la grammaire FB-LTAG.

Une méthode pour créer non pas manuellement mais automatiquement une grammaire généraliste est un objectif nouveau. La création manuelle implique un travail fastidieux et une faible adaptabilité, cette dernière étant le talon d'Achille des grammaires spécifiques à un domaine. La grammaire générée à partir de *treebanks* est d'un autre ressort. Elle offre une description d'un corpus. La technique que nous proposons permet l'extraction d'une grammaire adaptable, ciblant la possibilité de générer des phrases simples, à partir des données (exemples et annotations syntaxico-sémantique) de Verbnet (section 6). Nous montrons que cette grammaire permet de (re)-générer les exemples contenus dans Verbnet avec un score BLEU variant de 0.93 à 0.99 suivant la longueur des phrases (section 7).

2 Travaux connexes

Pour ce qui touche à la génération de grammaire noyau, la recherche oscille entre savoirs experts et restriction de l'espace syntaxico-sémantique. La création de grammaire exploitable en traitement automatique des langues est loin d'être un objectif neuf. Pour l'heure, il existe différentes méthodes : manuelle, en s'appuyant sur le travail de linguistes, automatique, en restreignant la tâche par le domaine (usage de données métier/d'un corpus annoté).
Le coût des savoirs experts reste notable, même dans un cadre où l'assistance par ordinateur est maximisée (Xia, 2001) : tout doit être pré-pensé, structuré "manuellement". Il en découle que la révision de la grammaire est difficile.
Se centrer sur un domaine particulier permet d'adapter aux données un algorithme d'étiquetage sémantique (Wong & Mooney, 2007) ou de développer des statistiques d'usage cadrées par l'expressivité et le vocabulaire propres au domaine (Gyawali, 2016).
Ciblant deux buts relativement proches, une autre méthode doit être mentionnée, une méthode stochastique. Celle-ci ne cible pas la création de grammaire noyau. Il s'agit de l'exploitation de *treebanks* : de corpus arborés (Chiang (2000); Xia (2001)). Cela permet la création d'une grammaire relativement couvrante, du moins vis-à-vis du corpus de référence. Par contre, nous dit Xia (2001), la consistance n'est pas garantie et la flexibilité reste pauvre. Pour un verbe utilisé d'une manière donnée, on en trouve douze répondant aux mêmes règles [douze est le nombre de verbes moyen par classe obtenu par Kipper *et al.* (2008)] : un corpus ne permet pas d'apprendre cela.

3 La grammaire, formalisme utilisé : FB-LTAG

Pour décrire la grammaire de l'anglais ou d'une autre langue de façon exploitable par un ordinateur, il est nécessaire d'utiliser un formalisme syntaxique précis et adéquat.

L'analyse syntaxique en TAG [1] se réalise en temps polynomial et permet notamment de décrire des contraintes entre constituants éloignés dans la phrase. L'expressivité est grande (classe de langage légèrement sensible au contexte). Enfin, il existe des outils pour la génération, comme GenI (Gardent & Kow, 2007a).

FB-LTAG [2] est une variante de TAG où chaque arbre TAG est associé avec un mot et décoré avec des structures de traits. Gardent & Kallmeyer (2003) montrent comment cette variante permet d'approcher la sémantique de manière compositionnelle en conjuguant sémantique et grammaire. C'est le formalisme que nous avons choisi.

Les arbres de la grammaire sont exploitables par le biais d'une sémantique. La phrase (1a) a pour sémantique plate la sémantique (1b).

(1) a. *A fire rage in the mountains.*
b. a(`w`) fire(`w`) rage(`e1`) Theme(`e1 w`) in(`e1 j`) the(`j`) mountains(`j`)

TABLE 1 – Sémantique plate

La phrase 1a utilise l'évènement `e1` *rage*. *Rage* est un évènement défini dans un lexique (cf figure 2) comme attendant un thème et comme pouvant s'ancrer dans un arbre où il sera spécifiable par une préposition. *In* est défini dans le lexique comme préposition pouvant mettre en relation x avec un évènement. Cet x ne pourra être que un groupe nominal (NP) car le groupe prépositionnel de l'arbre exploitable par *rage* a une définition stricte (voir Figure 1).

```
*ENTRY: rage
*CAT: v
*SEM: unaryRel[rel=rage,theta1=Theme]
*FAM: nOVPrepPN
*EX: {A_fire_raged_in_the_mountains.}
```

TABLE 2 – extrait lexical minimaliste : *rage*

Nous n'avons pas utilisé FB-LTAG directement. Nous l'avons fait générer par un compilateur de méta-grammaire, XMG, que nous allons présenter.

4 La méta-grammaire, XMG

Les méta-grammaires sont nées de la volonté d'offrir un langage simple d'utilisation, de pouvoir directement décrire ces notions syntaxiques de haut-niveau que sont les notions de sujet, d'objets, de verbe.

XMG est un terme faisant référence à la fois à un langage formel et à un compilateur éponyme permettant la génération de grammaires à partir de leur description (ou *métagrammaire*). Une définition formelle (syntaxe et sémantique) du langage XMG est donnée dans (Crabbé

1. TAG : Tree Adjoining Grammar, grammaire d'arbres adjoint
2. FB-LTAG : Feature-Based Lexicalized TAG, grammaire d'arbres adjoints lexicalisée basée sur des propriétés sémantiques

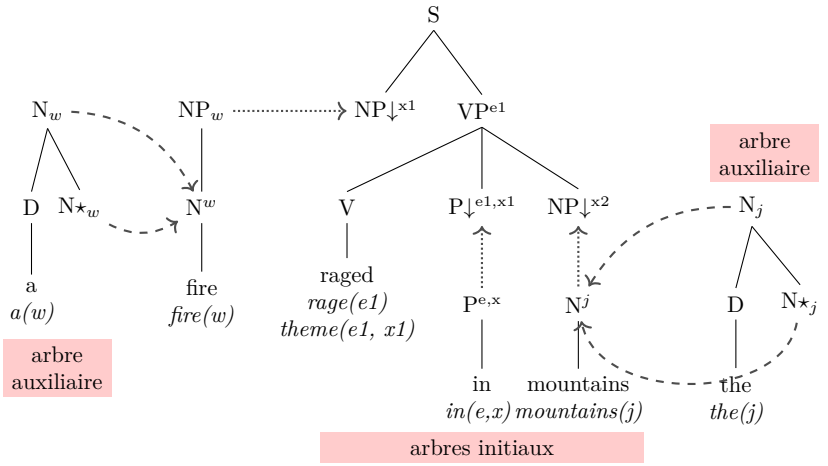

FIGURE 1 – "A fire raged in the mountains"

et al., 2013). Nous résumons ici brièvement les caractéristiques de ce formalisme utilisés dans le cadre de notre travail.

XMG permet de spécifier et de combiner des classes qui intuitivement, décrivent des fragments d'arbres, des fragments de représentations sémantiques et le partage de variables entre arbres et représentations sémantiques. Les fragments sont représentés sous forme de formules de logique de description d'arbres (Rogers & Vijay-Shanker, 1994). Les classes peuvent être combinées par héritage, conjonction et disjonction. Un système de couleur est également utilisable qui permet de contraindre l'unification de variables de nœuds d'arbre : un nœud rouge ne peut pas être identifié avec un autre nœud, un nœud noir peut être identifié avec un ou plusieurs nœuds blancs, et un nœud blanc doit être identifié avec un nœud noir. La Figure 2 illustre ces mécanismes avec de haut en bas, des fragments d'arbres, des fragments de représentations sémantiques et les équations permettant d'identifier des variables apparaissant dans les arbres d'une part et dans la sémantique d'autre part.

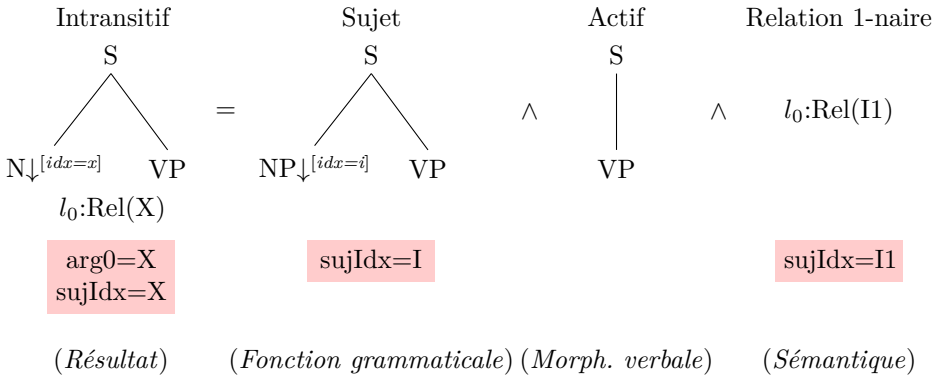

FIGURE 2 – Fragments XMG et leur combinaison

La description XMG d'une grammaire est souvent appelée une métagrammaire.

L'utilisation de XMG visait deux buts : réduire le coût de mise en œuvre de notre méthode,

par le biais de production intermédiaires intelligibles, et offrir un résultat aisé à maintenir : les fichiers XMG non compilés sont faciles à modifier.

XMG nous permet de gérer les informations sémantiques : la classe `unaryRel` correspond à la sémantique plate d'une relation unaire et permet le rapprochement des éléments lexicaux avec le contenu admis par des nœuds. Sa définition permet à un générateur le couplage de la sémantique avec les arbres compilés.

XMG nous a permis de décrire dynamiquement des relations unaires ou binaires, qu'un rôle attendant selon verbnet un objet de type '*adj*' pouvait accueillir aussi bien un participe passé (*vbd*) qu'un adjectif (*jj*), mais aussi qu'un adjectif pouvait modifier un nom pour devenir un groupe nominal, ce qui, après compilation, donne lieu à la création d'un arbre auxiliaire.

Nous avons utilisés les données VerbNet pour construire des classes XMG. Les notions de *modifiers* (un adjectif vient modifier le sens d'un nom) ont pu être gérées semi-automatiquement grâce à l'expressivité d'XMG, en partant du découpage des groupes syntaxiques puisés dans les phrases-exemple VerbNet et isolés informatiquement. La factorisation complètement automatisée serait, à ce niveau, trop complexe (erreurs d'étiquetage des composants de groupes syntaxiques). Faute de place, nous nous étendrons peu sur ce sujet ici, mais nous tenons à dire que la possibilité de décrire des alternatives entre fragments se prêtent fort bien à l'automatisation. La possibilité de nommer, d'architecturer les classes facilitent retours et adaptations.

5 VerbNet

Kipper *et al.* (2000) ont créé un lexique verbal exploitable pour les grammaires d'arbres adjoints. VerbNet 3.2, celui que nous utilisons, décrit 3695 verbes au sein d'un système de classes. Kipper *et al.* ont utilisé TAG pour construire leur lexique hiérarchisé au sein de classes.

Le système de classes de VerbNet est la structure accueillant les descriptions syntaxico-sémantique des verbes traités. Toute classe peut avoir 0 à n sous-classes. Une sous-classe est une classe qui hérite d'une autre classe.
Une classe ou sous-classe comprend un certain nombre de descriptions syntaxico-sémantiques. Nous allons voir comment ces descriptions sont structurées à l'aide d'identifiants syntaxiques et sémantiques, chaque identifiant étant doté de restrictions si nécessaire.

Chaque identifiant syntaxique (NP ou V, par exemple) est associé à un identifiant sémantique (tel NP est l'*Agent* dans la phrase, tel autre le *Thème*). Chaque identifiant, selon sa catégorie, peut être associé à des restrictions syntaxiques (ex : pluriel obligatoire) ou sémantiques (l'agent est *animate* ou *organization*).

Il y a 274 classes-racine dans VerbNet 3.2. Un verbe peut appartenir et en règle générale, appartient à plusieurs classes. La Figure 3 illustre les relations d'héritage de l'une des classes, *entity_specific_modes_being-47.2*.

Un verbe (tel *rage*) défini dans la classe *entity_specific_modes_being-47.2* a pour structures valides celles apportées par sa classe propre (*structure_a, structure_b, ...*). D'autres verbes

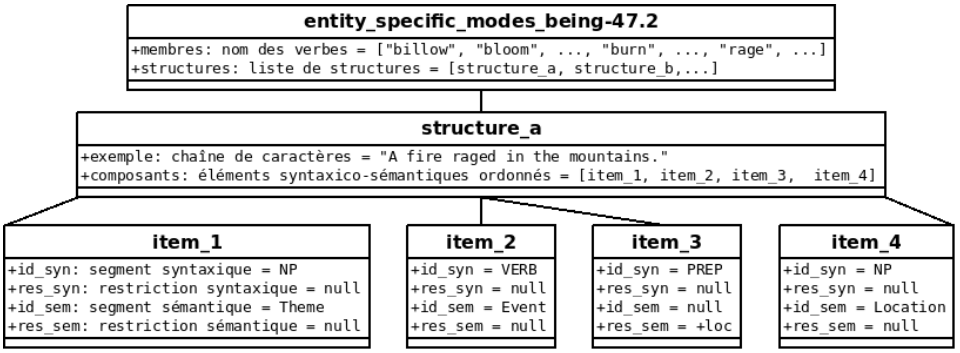

FIGURE 3 – illustration de la classe *entity_specific_modes_being-47.2*

(ex : *burn*) appartiennent à cette classe. Les structures décrites pour la classe sont aussi valides pour eux. Ces verbes peuvent être définis dans des sous-classes. Cela se produit quand un ou plusieurs verbes peuvent utiliser les règles de constructions de la classe-mère mais que d'autres structures spécifiquement exploitables par eux ont été référencées.
Un verbe (tel *report*) est défini dans la sous-classe *characterize-29.2-1-2* (en bas à droite Figure 4). Il a pour structures valides celles apportées par sa classe propre, mais il a aussi pour structure valides celles déclarées dans les classes dont il hérite : *characterize-29.2-1* et *characterize-29.2*. En bas à gauche Figure 4, *characterize-29.2-1-1*, qui existe en parallèle et ne le concerne pas, est également héritière de *characterize-29.2-1* et *characterize-29.2*.

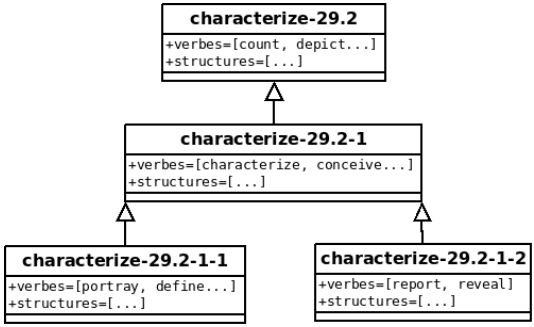

FIGURE 4 – Représentation de l'héritage sur *characterize-29.2*

L'extraction FB-LTAG, expliquée dans la partie suivante, est facilitée par cette organisation des données.

6 Extraction depuis VerbNet

Cette partie est consacrée à l'extraction automatique de la grammaire et du lexique verbal.

6.1 Vue d'ensemble

Nous avons vu que VerbNet décrit pour chaque classe les structures syntaxico-sémantiques utilisables par un nombre fini de verbes. VerbNet illustre chaque structure syntaxico-sémantique par un exemple (cf Figure 3).

Une structure syntaxico-sémantique est un quintuplet composé de :
- deux séries parallèles de segments syntaxiques,
- deux séries parallèles de segments sémantiques,
- un exemple.

La classe *entity_specific_modes_being-47.2* décrit la structure syntaxico-sémantique montrée dans la Table 3. Les séries de segments syntaxiques sont illustrés par les lignes a, les séries de segments sémantiques par les lignes b. Ces séries sont exemplifiées par la phrase c.

				références Figure 3			
				item_a	item_b	item_c	item_d
a	segments syntaxiques	a^1	identifiants	NP	Verb	PREP	NP
		a^2	restrictions	$\emptyset$	$\emptyset$	$\emptyset$	$\emptyset$
b	segments sémantiques	b^1	identifiants	Theme	Event	Prep	Location
		b^2	restrictions	$\emptyset$	$\emptyset$	$+loc$	$\emptyset$
c	exemple			A fire raged in the mountains.			

TABLE 3 – Structure syntaxico-sémantique adaptée à *A fire raged in the mountains.*

Toutes les informations syntaxiques nous ont été utiles, à la fois pour traiter les phrases exemple (segmentation en fonction des groupes attendus, caractérisation), à la fois pour introduire de fines et nécessaires contraintes syntaxiques. Les restrictions syntaxiques (il y en a 41 différentes) permettent de poser, par exemple, qu'un groupe nominal est systématiquement introduit par *how* pour une structure donnée. Manquant de place pour détailler les restrictions, nous renvoyons le lecteur à Kipper *et al.* (2008) pour plus d'informations.
Les restrictions sémantiques (ex : ligne b^2, Table 3) n'ont pas été exploitées. Elles seraient utiles, pour spécifier par exemple lexicalement les adverbes et restreindre les possibles en génération. Ce choix n'est lié qu'à des contraintes de temps. Les identifiants sémantiques nous ont permis de déterminer les rôles sémantiques des arguments syntaxiques et des modifieurs (sujet, COD, compléments circonstanciels...).

Les composants, segments syntaxico-sémantiques, parmi lesquels les items *item_a*, *_b*, *_c* et *_d* introduits Figure 3, sont au nombre de 192. Avec leur aide, nous extrayons de l'ensemble des classes fournies par VerbNet une grammaire d'arbres adjoints noyau reflétant les structures verbales décrites dans cette ressource. La procédure d'extraction se fait selon les étapes suivantes :

- segmentation syntaxico-lexicale : chaque exemple est segmenté en fragments ; le couplage des informations initiales (lignes a, b et c, Table 4) à ces fragments (d) est effectué en phase avec les informations fournies par une table de correspondance (ligne e, Table 4).
- construction des fragments d'arbre (ligne f) et sémantiques (ligne g) qui permettront de créer une grammaire sous forme de classes (méta-grammaire) et de tester la grammaire

noyau,
— compilation de la métagrammaire.

Pour évaluer la grammaire résultant de l'extraction, nous créons à partir de chaque exemple Verbnet et des rôles qui lui sont associés, une sémantique plate reflétant la sémantique de l'exemple (e.g., (1b) pour (1a)). Nous donnons ensuite en entrée au générateur GenI (Gardent & Kow, 2007b), ces représentations sémantiques couplées avec la grammaire lexicalisée extraite par notre procédure d'extraction et nous comparons la phrase générée par GenI avec l'exemple VerbNet initial en utilisant la métrique BLEU (Papineni *et al.*, 2002) qui positionne l'évaluation de qualité entre 0 et 1. Un score de 1 indique que la phrase générée est identique à l'exemple VerbNet initial.

6.2 Table de correspondance

La table de correspondance nous permet de définir les clés support à la factorisation des arbres. Elle permet la génération et l'exploitation d'un fragment d'arbre associé à chaque composant décrit par VerbNet.

Les données de correspondance répondent à trois besoins :
① nom de clé à positionner pour construire le fragment d'arbre comprenant le segment,
② segment à traiter par le biais d'un rôle ou bien sous forme de co-ancre,
③ segment impliquant la présence d'un élément le préfixant ou le suffixant.

Le composant *item_a* (Figure 3 et Table 3) est spécifié par VerbNet comme étant un `NP`, identifié sémantiquement comme `Theme`, et dépourvu de restriction syntaxique. On y associe une clé (`n` pour notre propre table de correspondance). L'arbre initial exploitant ce fragment verra son identifiant XMG bâti avec cette information (`n0VPrepPN` pour l'arbre généré à partir de la structure de *A fire raged in the mountains.*). Cela permet la gestion de la factorisation des arbres, donc la réexploitation des fragments et des arbres initiaux.

Le traitement des besoins ① et ② seront explicités dans la partie suivante, qui traite le processus d'extraction.

Le besoin ③, très spécifique, est lié aux restrictions syntaxiques. Par exemple, un segment syntaxique *NP* associé à l'information sémantique *Value* avec la restriction sémantique *+how_extract* donnera naissance à un nœud nommé *H* coancré par *how*, de catégorie *WRB*[3]. Cela implique la recherche du mot *how* lors de la fragmentation de la phrase exemple, la génération d'une co-ancre *how* dans l'arbre initial créé à partir des informations : sa production au sein de la méta-grammaire contenante.
La table de correspondance fournit l'appui nécessaire aux traitements.

86% des types de segment syntaxico-sémantique traités donneront lieu à la recherche d'un groupe nominal, d'un adverbe, d'un verbe ou encore d'un adjectif et à la création d'un nœud TAG correspondant. Pour eux, la table de correspondance suffit.
Le traitement est spécifique pour 27 types de nœuds sur les 192 potentiellement gérables, ce qui signifie la production de traits sémantiques adaptés lors de la génération des fragments.

3. WRB : catégorie de *how* au sein du jeu d'étiquettes du Penn Treebank Project, qui est le jeu que nous avons utilisé dans le cadre de cette expérience pour catégoriser les segments et leurs composants.

Le programme implémenté dans le cadre de l'expérience vient spécifiquement contraindre certains fragments.

6.3 Exemple d'extraction

La Table 4 illustre le processus d'extraction. Les trois premières lignes montrent la segmentation de l'exemple *A fire raged in the mountains* à partir des informations sémantiques (a), syntaxiques (b et c) en mots (d), ce guidés par les informations de correspondance (e). Ces étapes permettent la récupération de la structure XMG initiale pour le verbe. La ligne f donne le nom des classes XMG correspondantes et la ligne g montre la sémantique plate qui leur est associée.

a. *Structure sémantique*	Theme	Event		Location
b. *Structure syntaxique*	NP	VERB	PREP	NP
c. *Restrictions syntaxiques*	∅	∅	∅	∅
d. *Exemple*	A fire	raged	in	the mountains
e. *Correspondance*	N *rôle*	V	P *rôle*	N *rôle*
	↓	↓	↓	↓
f. *Classes XMG*	NObjectB-11	n0VPrepPN	redprainbs	PrepPNObjectA-12
g. *Sémantique plate*	i :a_fire(w)	r :rage(e1) r :Theme(e1 w)	m :in(e1 j)	z :the_mountains(j)

Table 4 – Extraction d'un arbre intégrant un modificateur

La Figure 5 montre les fragments XMG résultant de notre procédure d'extraction. Cette procédure spécifie en outre que le fragment *NObjectB-11* (i) importe le fragment *Fragment Verbal importé* et (ii) est combiné par conjonction avec le fragment *PrepPNObjectA-12*. La compilation XMG produit l'arbre TAG n0VPrepPN (cf. Figure 1) qui sera associé dans le lexique de génération avec le verbe *to rage* ainsi qu'avec tous les verbes de la classe *entity specific modes being*[4].

Nous aboutissons à la génération automatique des arbres TAG, comme celui de n0VPrepPN, créé pour l'exemple présenté Table 4 de la classe *entity specific modes being*.
La sémantique plate permet de tester avec un réalisateur de surface que le texte est correctement généré à partir des arbres. Elle est formée en parallèle de la grammaire lexicalisée des verbes. Cela nous amène presque à la fin de cet article, aux tests, et donc aux résultats.

7 Résultats

VerbNet nous fournit 1415 exemples pour couvrir ses verbes. Certains exemples (91 sur 1415, soit 6%) servent à plusieurs structures. '*The matter seems in dispute.*' sert à couvrir une structure de la classe *seem-109.xml* et une autre de la classe *become-109.1.xml*. Dans la première il illustre une structure thème/verbe/attribut sentencieux[5], dans la seconde, une

4. *entity specific modes being* : cette classe regroupe des verbes permettant de spécifier une façon spécifique d'être pour des entités (ex : *burn*, *fizz*, *flower*)...

5. Extrait de VerbNet, *seem-109.xml* :
`example "The matter seems in dispute." → syntax Theme V {in} Attribute<-sentential>`

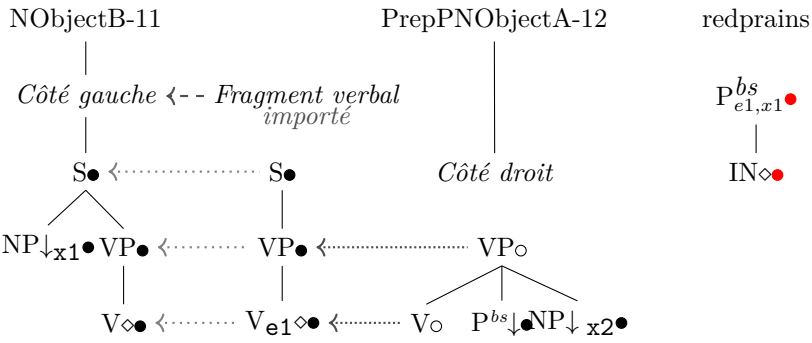

FIGURE 5 – composition de n0VPrepPN et fragment pour préposition

structure thème/verbe/résultat. Les exemples réutilisés peuvent l'être jusqu'à cinq fois.
De la même manière, inversement, des structures peuvent se voir illustrées par plusieurs exemples. Patient/verbe/with/co-patient est dans ce cas, elle a quatre exemples issus de quatre classes pour la démontrer :

— *The eggs mixed with the cream.*
— *My computer connected to his computer.*
— *Ice cream integrates with desserts.*
— *Company A associated with Company B.*

Nous avons 720 structures possibles, avec 514 qui sont illustrées par un seul et même exemple, les 206 autres qui sont dotées en moyenne de 4,78 exemples.
VerbNet est relativement riche, avec des phrase canoniques relativement courtes (5.93 mots en moyenne).

Sur notre corpus constitués des 1415 phrases-exemple de VerbNet, nous générons 1711 phrase à partir des représentations sémantiques plates dérivées des données VerbNet, de la grammaire extraite et du générateur GenI. Nous obtenons un score BLEU-4 [6] (utilisé dans 90% des cas, composés en moyenne de 6.3 mots) de 0.93, un BLEU-3 de 0.96 pour les 7% de trigrammes, et un BLEU-2 de 0.99 pour les 3% de bigrammes).

Si le traitement des phrases simples (ne contenant pas de subordonnées) est opérationnel, le traitement des subordonnées n'est pas pleinement fait. Les données de VerbNet ont permis d'y trouver des incohérences (par exemple, pour une structure déclarée NP V NP LEX$^{=up}$, l'exemple VerbNet fournissait un NP après le mot *up* syntaxiquement imposé comme clôturant la phrase).

Des corrections mineures ont donc été requises. Les caractéristiques des subordonnées sont fournies par VerbNet, elles ont donc été implémentées en tant que contraintes sur le groupe nominal décrit. Non seulement le fragment phrasique de la subordonnée n'a pas été créé pour que des éléments puissent s'y ancrer, mais, de fait, les incohérences n'ont pas été relevées. Concrètement, un arbre final existe attendant un fragment avec des caractéristiques propres

6. BLEU : Bilingual Evaluation Understudy : mesure de la proximité entre deux entrées, typiquement entre une sortie machine et une phrase en langue naturelle, BLEU-4 s'appuie sur des quadrigrammes (comparant les mots par succession de quatre), et n'est donc pas exploitable pour des phrases de trois mots où BLEU-3 peut être utilisé...

à la subordonnée attendue (ex : qu'elle doit être introduite par *that* et être à tel temps par rapport à la principale) ; on vérifie qu'elle peut se réaliser (si toute fois les fragments permettaient d'ancrer la partie adjointe) : par le biais du lexique, on fournit un groupe ayant les bonnes caractéristiques. Rien encore n'est prêt pour que ce groupe puisse être réalisé par le biais de la sémantique. Il y a quelques erreurs dans VerbNet. Quand un élément est fourni dans une structure comme élément lexical coancré et à la fois est indiqué par des informations syntaxiques que nous traitons comme des traits, cela nous donne une surgénération aberrante du type '*he asked what whether to go*' ou encore '*he asked her to to do*'.

La surgénération non souhaitable représente approximativement 7% des hypothèses produites, ce qui nous laisse un corpus correct de quasiment 1600 phrases pour nos 1415 références, et explique la hauteur du score BLEU.

Un faible pourcentage des surgénérations est lié au non usage de traits dans le cadre sémantique. Nous créons *ad hoc* le lexique, et '*he converted*' se voit généré aussi bien que '*he was converting*' pour une même sémantique alors que '*he was converting*' ne fait pas partie du corpus de référence. Les temps auraient du être fournis dans des traits au niveau sémantique, trouver leurs formes dans un lexique morphologique, et non figurer comme variation lexicale d'un même lemme verbal dans le lexique dédié.

Un certain nombre de phrases ne figurant pas dans le référentiel ont été générées tout en étant correctes pour la sémantique donnée. Par exemple, '*ellen told helen the situation*', '*ellen told the situation to helen*' ont été offertes en variation de '*ellen told helen about the situation*'.

Nous aurions du couvrir 100% du corpus. Il y a trois phrases que nous ne parvenons pas à regénérer telles qu'elles auraient du être (la diversité sémantique permet de générer une phrase autre, mais pas celle de départ). Comme elles sont peu nombreuses, nous pouvons nous y attarder.

- '*she says ' enchilada ' with a proper mexican accent*' → mauvaise lexicalisation, l'arbre est correct (perte d'une cote dans l'information générée).
- '*i broke the twig and the branch apart*' → l'arbre est correct, *the twig and the branch apart* forment ici un groupe nominal complet structurellement, notre procédure de séparation des groupes (nominaux/verbaux/prépositionnels) adaptée à VerbNet a échoué sur cet exemple, ce qui s'est traduit par une mauvaise lexicalisation puis une mauvaise regénération. [7].
- '*on his finger there sparkled a magnificent diamond*' → mauvaise gestion d'un sujet anonyme après un groupe prépositionnel (les sujets anonymes ont un traitement spécifique).

Nous profitons du deuxième exemple pour mettre à jour une faille expérimentale : les traits ont été sculptés sur mesure pour la phrase canonique type (sujet verbe complément(s)), le cas du sujet anonyme a été pris en charge sauf pour l'exemple pré-cité. Néanmoins, il serait nécessaire de pousser l'expérience à une véritable interprétation de ce qui est sujet : « All through the mountains raged a fire. » est un type de phrase représentant 0.8% du corpus. Sur ce 0.8% du corpus, l'accord sujet/verbe ne se ferait pas correctement : les traits ont été portés pour se synchroniser du premier groupe au second dans le cadre d'une structure sujet verbe suivis de compléments très largement dominante dans les données VerbNet.

7. lexicalisation : dans le cadre de cette expérience, pour les tests, les groupes nominaux ont été traité comme un tout, l'usage segmenté de leur contenu fait partie d'une autre expérience

Au final, à partir des structures canoniques de VerbNet, nous obtenons en génération 172 arbres initiaux composés de deux à trois fragments (côté gauche, côté droit, noyau verbal importé) couvrant, plus ou moins bien (le traitement des subordonnées étant à finaliser).

Pour ce qui est des fragments générés, les côtés gauches sont au nombre de 8, les côtés droits 123. Il faut ajouter aux côtés gauches le sujet anonyme, exploité par sept des 172 arbres initiaux au-delà de l'arbre initial réflexif évoqué Table 6. La classe XMG sujet anonyme a été fournie manuellement, comme les deux noyaux atomiques verbaux (nombre libre ou pluriel contraint).

Les fragments générés (voir Table 6) reposent sur l'exploitation de fragments produits manuellement (voir Table 5).

arbres initiaux	**exemple**
noyau verbal	*There* ***happened*** *an accident*
sujet anonyme	***There*** *happened an accident*
binaire : attribut d'objet	*thief* ***of*** *diamonds, the shelf* ***of*** *drinks*

Table 5 – Arbres initiaux fournis

arbres initiaux finaux	**nb**	**exemple**
réflexif	1	*It rains.*
unaire	23	*Susan was chitchatting.* ou *There happened an accident.*
binaire	73	*Carmen bought a dress.*
ternaire	70	*Ellen told Helen to come.*
quaternaire	5	*Carol cut the envelope open with the knife.*
autres initiaux		
unaire	11	*Carol*, *nice*, *the*, *really* ...
unaire ou binaire	2	prépositions : *to*, *in*, *from*, adverbes :*down*

Table 6 – Arbres initiaux générés

8 Conclusion

L'utilisation des ressources linguistiques permet, sans expertise linguistique propre à l'utilisateur, de générer un lexique verbal adossé à la grammaire TAG qui lui correspond. Plusieurs sources d'informations ont du être couplées pour valider les formes que peut prendre un verbe et pour définir finement les composants des groupes syntaxiques (pour information, il s'agissait de NLTK et Wordnet, que nous ne présentons pas ici car non centraux dans la méthode).
Nous pouvons attendre de notre système un taux catégoriel de phrases correctes, en l'état, à 92% (les 8% manquants étant les cas non générés, la surgénération et les soucis de traits).

Ce système est objectivement fonctionnel, exploitable.
L'usage de TAG permettant l'expansion de n'importe quel groupe, son formalisme permettant la manipulation, la continuation de la recherche est plus qu'envisageable.

Il serait nécessaire de gérer les traits sémantiques (eux aussi fournis dans verbnet) pour cadrer par exemple un *Agent*, le cas échéant, sur un trait type *animate* ou *organization* (ce qui

signifie de manière relativement claire « l'agent doit être animé ou être une organisation »). Couplés à la notion de rôles, ces informations parferaient le modèle et seraient utiles pour adosser à notre humble grammaire formes transitives et interrogatives. Sachant comment on pose une question dans une langue donnée, il est possible, en remplaçant un rôle par une préposition introductive, de donner une forme cohérente à une interrogation.

Au-delà des bases verbales (formes morphologiques et lexique liant verbes et arbres), de la grammaire produite, une architecture devrait être pensée pour faciliter l'emploi d'une production intermédiaire issue d'un tel processus car la production intermédiaire est elle-même très riche (informations détaillées touchant aux structures et rôles).

Pour en revenir à notre choix - profiter des larges ressources commençant à s'accumuler en TAL, nous avançons ce diagnostic qui peut sembler aller de soit après la lecture de ce bref article, mais que l'état de l'art ne montre pas être si courante. Le traitement de la langue est un domaine trop vaste pour pouvoir le révolutionner seuls, il est productif de maximiser l'usage de briques conçues par la communauté, dans une perspective visant parfois la "simple" meilleure compréhension fondamentale de la langue.

Références

Cahill A. & Van Genabith J. (2006). Robust pcfg-based generation using automatically acquired lfg approximations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, p. 1033–1040 : Association for Computational Linguistics.

Candito M.-H. (1996). A principle-based hierarchical representation of ltags. In *Proceedings of the 16th conference on Computational linguistics-Volume 1*, p. 194–199 : Association for Computational Linguistics.

Candito M.-H. (1998). Building parallel ltag for french and italian. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, p. 211–217 : Association for Computational Linguistics.

Carroll J. & Oepen S. (2005). High efficiency realization for a wide-coverage unification grammar. In *International Conference on Natural Language Processing*, p. 165–176 : Springer.

Chiang D. (2000). Statistical parsing with an automatically-extracted tree adjoining grammar. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, p. 456–463 : Association for Computational Linguistics.

Crabbé B., Duchier D., Gardent C., Le Roux J. & Parmentier Y. (2013). Xmg : extensible metagrammar. *Computational Linguistics*, **39**(3), 591–629.

Gardent C. (2006). Intégration d'une dimension sémantique dans les grammaires d'arbres adjoints. In *Actes de la conférence TALN 2006*, p. 149–158.

Gardent C. & Kallmeyer L. (2003). Semantic construction in feature-based TAG. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*, p. 123–130 : Association for Computational Linguistics.

Gardent C. & Kow E. (2007a). GenI, un réalisateur basé sur une grammaire réversible. In *14e conférence pour le Traitement Automatique des Langues Naturelles-TALN 2007*, p. 10–p.

Gardent C. & Kow E. (2007b). A symbolic approach to near-deterministic surface realisation using tree adjoining grammar. In *45th Annual Meeting of the Association for Computational Linguistics-ACL 2007*, p. 328–335 : Association for Computational Linguistics.

Gyawali B. (2016). *Surface Realisation from Knowledge Bases*. PhD thesis, Universite de Lorraine.

Kipper K., Dang H. T., Palmer M. & others (2000). Class-based construction of a verb lexicon. *AAAI/IAAI*, **691**, 696.

Kipper K., Korhonen A., Ryant N. & Palmer M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation*, **42**(1), 21–40.

Montague R. (1973). The proper treatment of quantification in ordinary English. In *Approaches to natural language*, p. 221–242. Springer.

Narayan S. & Gardent C. (2012). Structure-driven lexicalist generation. In *24th International Conference in Computational Linguistics (COLING)*, p. 100–113.

Papineni K., Roukos S., Ward T. & Zhu W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, p. 311–318 : Association for Computational Linguistics.

Rogers J. & Vijay-Shanker K. (1994). Obtening Trees from Their Descriptions : An Application to Tree-Adjoining Grammars. *Computational intelligence*, **10**(4), 401–421.

Vijay-Shanker K. & Joshi A. K. (1988). Feature structures based tree adjoining grammars. In *Proceedings of the 12th conference on Computational linguistics-Volume 2*, p. 714–719 : Association for Computational Linguistics.

Wong Y. W. & Mooney R. J. (2007). Learning synchronous grammars for semantic parsing with lambda calculus. In *Annual Meeting-Association for computational Linguistics*, volume 1, p. 960.

Xia F. (2001). *Automatic grammar generation from two different perspectives*. phdthesis, University of pennsylvania.