Normalisation de la représentation des lexiques syntaxiques arabes pour les formalismes d'unification

Noureddine Loukil [1], Kais Haddar [2], Abdelmajid BenHamadou [1]
Laboratoire de Recherche en Informatique et Multimédia, Sfax, Tunisie

Abstract

La structure et la représentation des ressources lexicales sont nécessaires pour élaborer des lexiques réutilisables à large couverture. Dans ce papier, nous étudions cette nécessité pour les lexiques syntaxiques pour la langue arabe. Pour cela, nous avons mis en avant les spécificités des lexiques HPSG et LTAG et dégager une structuration et une représentation communes. Nous nous sommes, par la suite, basé sur le Lexical Markup Framework (LMF), une récente proposition de normalisation ISO, pour coder le lexique. LMF semble pouvoir supporter ce rôle grâce à sa généralité et à son extension syntaxique.

Keywords : Lexique syntaxique, HPSG, LTAG, Représentation lexicale, Lexical Markup Framework.

1. Introduction

La plupart des nouveaux formalismes linguistiques utilisés dans le traitement automatique de la langue naturelle expriment une tendance à stocker le maximum de connaissances linguistiques dans le lexique et à réduire la grammaire en un ensemble de principes et/ou de schémas universels qu'il suffit d'étendre et/ou de personnaliser pour prendre en compte des particularités de chaque langue. De plus, malgré leur diversité théorique, ces formalismes proposent tous l'idée de dissocier le lexique de la grammaire d'une manière laissant espérer le développement des ressources lexicales indépendamment du formalisme utilisé.

Les formalismes d'unification comme les grammaires d'arbres adjoints, TAG (Kroch & Joshi, 1985) et les grammaires syntagmatiques guidées par les têtes, HPSG (Pollard & Sag, 1994) adoptent cette nouvelle vision et puisent l'essentiel de l'information linguistique dans un lexique qui fournit une description morphologique, syntaxique et sémantique pour chaque mot.

Dans le même sens, plusieurs travaux récents ont établi des bases théoriques et pratiques pour dissocier le processus de développement de la grammaire de celui du lexique. Par exemple, (Candito, 1996) définit une méthodologie pour la création des grammaires TAG d'une manière abstraite en définissant la nouvelle abstraction de métagrammaire. Aussi, (Gaiffe & al., 2002) propose un compilateur qui servira pour générer automatiquement le lexique. L'entrée de ce compilateur est une métagrammaire écrite par le linguiste dans un langage de haut niveau. Quant à la sortie, c'est un ensemble d'arbres élémentaires constituant le lexique et la grammaire LTAG générés.

Le développement du lexique syntaxique pose des problèmes concernant la structuration et la représentation qu'il faut adopter (Francopoulo, 2003). En effet, la diversité des représentations rend difficile la diffusion et l'échange entre les différentes sous communautés et empêche, en particulier, la construction de lexiques à large couverture. Le besoin d'une norme est évident surtout pour les communautés linguistiques qui sont dépourvues de moyens intéressants. Ainsi,

[1] {noureddine.loukil, abdelmajid.benhamadou}@isimsf.rnu.tn
[2] kais.haddar@fss.rnu.tn

l'adoption d'une telle norme par tout le monde servira d'un axe pivot pour l'interopérabilité des applications et l'interchangeabilité des ressources lexicales. La communauté de TAL arabe incarne cette situation. En fait, plusieurs travaux ont essayé de créer des ressources lexicales syntaxiques. Parmi ces travaux, nous citons (Bahou & al., 2003) et (Elleuch & al., 2002) qui focalisent sur la représentation des verbes arabes dans le lexique HPSG. (Dichy, 2003) définit les traits associés aux entrées lexicales dans un lexique morphosyntaxique pour la langue arabe et propose une ressource lexicale payante dans le cadre du projet MBC-DIINAR. Tous ces travaux partagent le but de créer d'une ressource lexicale syntaxique pour la langue arabe mais avec deux inconvénients majeurs : (i) les formats des lexiques sont propriétaires rendant difficile voire impossible l'utilisation de la ressource dans d'autres projets et leur fusion avec d'autres ressources. (ii) Ces lexiques sont intégrés dans des applications et ne présentent pas une interface logique ou logicielle facilitant leur utilisation.

Plusieurs travaux récents se sont intéressés à la question de normalisation au niveau de la représentation et de la structuration. (Loukil, 2006) et (Fehri & al., 2006) proposent une représentation normalisée pour le lexique HPSG arabisé, (Akrout, 2005) propose une représentation normalisée pour le lexique morphologique arabe. Cet effort n'est pas conduit loin d'autres projets ayant le même objectif. Parmi ces projets, nous citons :

– Le projet SYNTAX [1] qui intègre, entre autres, LMF et un registre des catégories de données utilisé dans les ressources lexicales.
– Le projet MOSAIQUE [2] qui vise la création d'un modèle syntaxique de haut niveau pour les lexiques syntaxiques.
– Le projet LEXSYNT [3] qui se propose plutôt comme un développement vertical pour la conception d'un lexique syntaxique sémantique de référence pour la langue française.

Ce papier introduit, tout d'abord, l'ensemble des concepts impliqués dans la conception et la structuration des lexiques pour les formalismes HPSG et TAG. Ensuite, nous essayons de dégager les points communs et les divergences dans l'optique d'une conciliation menant à adopter un modèle commun qui accepte les particularités de chacun. Nous proposons, par la suite, une méthodologie pour la conversion de lexiques HPSG et LTAG vers un lexique LMF normalisé. Puis, nous discutons les problèmes qui doivent être traités pour garantir l'utilisation de cette proposition.

2. Anatomie des lexiques utilisés par les grammaires d'unification

Les lexiques utilisés par les applications de TALN sont généralement divisés en deux types : les lexiques morphosyntaxiques et les lexiques syntaxiques. Les lexiques syntaxiques ou lexiques-grammaires fournissent une description détaillée des comportements syntaxiques et des figures de valence pour chaque entrée lexicale et proposent généralement des informations concernant le nombre et nature des arguments i.g., complément, infinitif ou groupe nominal humain, les prépositions appropriées, les transformations acceptées i.g., passif, construction croisée, effacement d'un argument et la résolution de co-références i.g., entre le sujet d'une infinitive et un argument du verbe principal. Dans les sous sections suivantes, nous présentons la structuration des lexiques utilisés par les formalismes HPSG et LTAG et nous discutons des relations conceptuelles entre les deux.

2.1. Anatomie du lexique HPSG

Un lexique HPSG est composé d'un ensemble d'entrées lexicales. Chaque entrée est exprimée par une structure attribut valeur. La figure 1 représente la structure attribut valeur du verbe "*passer*". Une spécification du type est incluse dans chaque entrée permettant de surcharger l'entrée par un ensemble de contraintes préalablement défini. Le type de la structure de la figure 1 est *verbe transitif*.

Les types qui sont eux-mêmes exprimés sous la forme de structures attribut valeur sont classés dans une hiérarchie dans la quelle chaque noeud hérite des propriétés des noeuds ancêtres.

[1] http ://syntax.inist.fr
[2] http ://mosaique.labri.fr
[3] http ://lexsynt.inria.fr

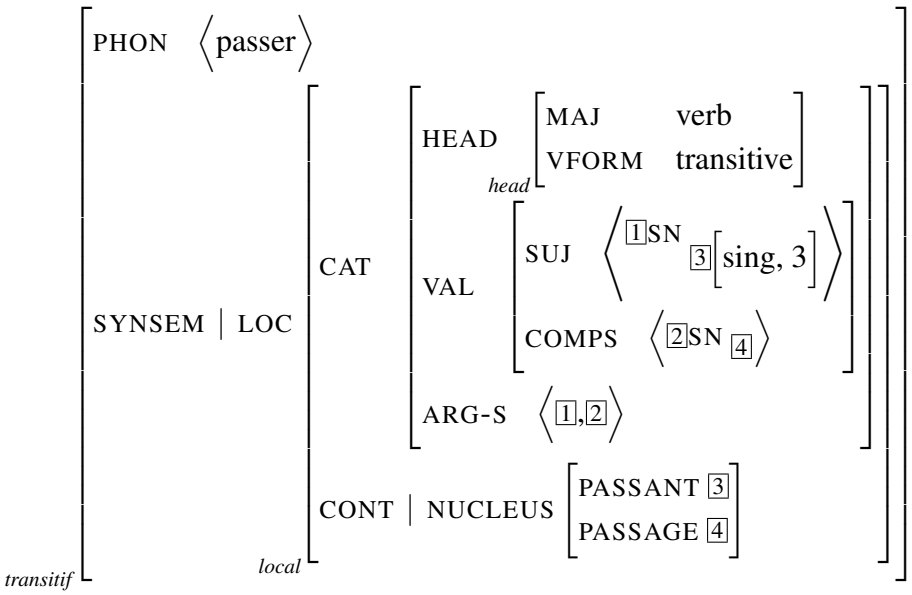

FIG. *1. Entrée lexicale du verbe* passer

Un morçeau de la hiérarchie de types est donné dans la figure 2. Ce mécanisme de factorisation des propriétés dans les types permet de libérer le lexique d'une redondance naturelle. De ce fait, la construction d'un nouveau lexique passe inévitablement par l'établissement d'une hiérarchie de types qui correspond à l'ensemble des objets linguistiques manipulés. Un lexique HPSG étant supposé contenir toutes les formes fléchies, ie un lexique en extension, les régularités flexionnelles et dérivationnelles entre les différentes entrées peuvent être décrites par des relations un à un reliant les entrées entre eux. Ces relations sont modélisées par les règles lexicales qui sont considérées de point de vue théorique comme des relations reliants les différentes entrées lexicales. En résumé, on peut dire qu'un lexique se compose de la signature (la hiérarchie des types) imposée sur le domaine linguistique et d'un ensemble de contraintes formulées sur les entrées lexicales. De plus, un ensemble de relations entre les entrées lexicales est formulé par des règles lexicales. Donc, on peut considérer trois composantes : la hiérarchie de types, les entrées lexicales et les régles lexicales.

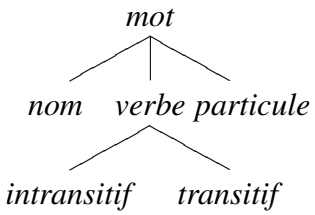

FIG. *2. Hiérarchie de types*

2.2. Anatomie du lexique LTAG et comparaison avec HPSG

Les entrées lexicales dans un lexique LTAG sont des arbres élémentaires, unités de base du formalisme. L'absence dans la théorie d'une méthode de factorisation lexicale a poussé les constructeurs de grammaires LTAG à doter le formalisme d'outils supplémentaires pour rendre possible la construction d'une ressource à large couverture. La construction d'un lexique LTAG revient en théorie à établir la liste des arbres élémentaires qui représentent les structures de description de base du formalisme. Un arbre est attribué pour chaque lexéme et pour chaque flexion. Cette méthode de description plate engendre une gigantesque quantité d'entités distinctes avec une redondance évidente. Cela crée un probléme de maintenance de la ressource d'un point de vue pratique et pose aussi des questions de cohérence théorique.

L'élégance théorique des hiérarchies de type faisant partie intégrante dans le formalisme HPSG s'est vite trouvé un similaire dans le formalisme LTAG. En fait, (Candito, 1996) a introduit une abstraction touchant les arbres élémentaires : il suffit de les sous-spécifier pour généraliser la description et avoir, ainsi, une description commune à plusieurs entrées lexicales. De cette façon, un arbre élémentaire non instancié (schème) peut jouer le rôle du type en HPSG.

Les schèmes proposent un typage correspondant aux types feuilles de la hiérarchie de types en HPSG. Les nœuds supérieurs de la hiérarchie correspondent à des super types et fournissent une description de moins en moins spécifiée donc de plus en plus générale. (Abeillé, 1991) a établi un deuxiéme niveau d'abstraction en introduisant le notion de « famille de schèmes ». Une famille est un ensemble de schémes partageant la même structure prédicative (par exemple : n0V, n0Vn1). Cette abstraction supplémentaire complète le concept général de la hiérarchie de types. En fait, les schèmes représentent les types feuilles directement instanciables tandis que les familles de schèmes ou « classes » représentent les types parents ou les super types. La figure 3 dresse les correspondances de point de vue concept entre le lexique HGSP et TAG.

Concept	HPSG	LTAG
Entrées lexicales	Structures de traits	Arbres élémentaires
Factorisation lexicale	Hiérarchie de types	Schèmes
Relations lexicales	Règles lexicales	-

FIG. *3. Correspondances conceptuelles entre les lexiques HPSG et LTAG*

Nous signalons qu'on a omis la correspondance LTAG des règles lexicales. En effet, nous ne traitons pas ce mécanisme dans LTAG d'autant plus que les nouveaux travaux tendent à ignorer le rôle productif de telles règles dans ce formalisme.

3. Présentation du Lexical Markup Framework

LMF (Francopoulo & George, 2005) est basé sur une organisation sémasiologique des entrées lexicales. LMF est composé d'un méta modéle de base, d'une structure squelette décrivant la hiérarchie des informations incluses dans une entrée lexicale. De plus, LMF spécifie les catégories spécifiques de données pour la variété des types de ressources et les contraintes gouvernant la relation de ces catégories de données au meta-modéle et à ses extensions. LMF offre aussi des procédures standard pour exprimer les catégories de données et les objets informationnels liés sous forme d'éléments XML et attributs.

3.1. Le modéle de base de LMF

Le modèle de base de LMF, présenté à la figure 4, est une structure hiérarchique composée des composants suivants : (1) Le composant «Lexical Database» rassemblant toutes les informations liées à un lexique donné, le composant «Global Information» rassemblant des méta-données (version, contributeurs, mises à jour,...). (2) Le composant «Lexical Entry» représentant une unité lexicale élémentaire dans le lexique. (3) Le composant «Form» offrant une représentation des propriétés phonologiques, morphologiques et flexionnelles pour les différentes réalisations morphologiques d'une entrée lexicale. Et (4) le composant «Sense» qui fournit une sémantique pour l'entrée lexicale et qui peut être divisé en des sous sens.

3.2. L'extension syntaxique de LMF

Le diagramme de l'extension syntaxique donnée à la figure 5 est basé autour du composant «comportement syntaxique». Un comportement syntaxique est un patron de construction syntaxique qui peut être utilisé par plusieurs entrées lexicales permettant ainsi de factoriser le même comportement syntaxique utilisé par plusieurs entrées lexicales et d'éviter la redondance. Un comportement syntaxique est décrit par l'ensemble des constructions syntaxiques permises éventuellement groupées dans des sous-ensembles de significations sémantiques disjointes. Un «Frame» représente synthétiquement un ensemble de structures syntaxiques possibles associées à un prédicat.

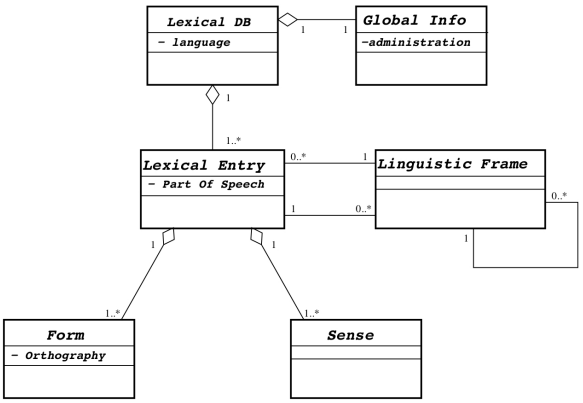

FIG. *4. Le modèle de base de LMF*

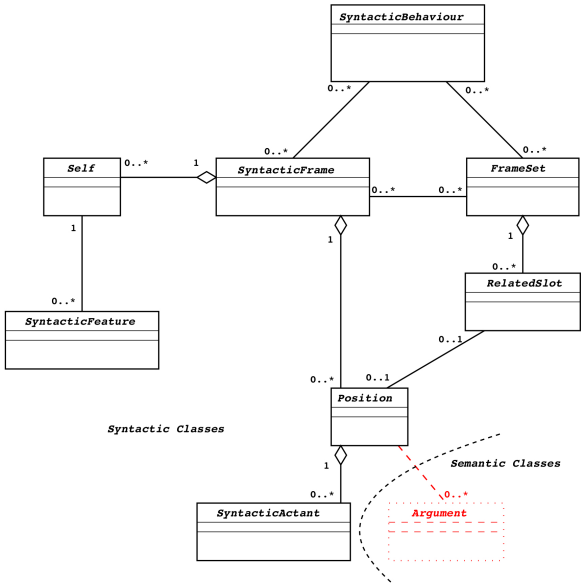

FIG. *5. Extension syntaxique de LMF*

Pratiquement, cela revient à une construction syntaxique particulière réalisée à l'aide d'un ensemble de compléments ou positions. Dans un «Frame», les réalisations combinatoires de ces positions mènent à des instanciations possibles de surface de ce «Frame» et donc à des phrases syntaxiquement correctes. En d'autres mots, le «Frame» peut être connu comme un patron valenciel fournissant une spécification de l'ordre et de la nature des compléments permis pour la formation d'une phrase acceptée. De ce fait, il faut compter plusieurs «Frame» pour une seule entrée lexicale. Chacun propose une ou plusieurs positions nécessaires ou optionnelles. Chaque position propose à son tour des réalisations possibles avec leurs descriptions morphosyntaxiques modélisées par le composant «SyntacticActant». Le composant «Self» décrit les propriétés morphosyntaxiques de l'entrée lexicale en question.

4. Représentation automatique des lexiques syntaxiques avec LMF

Les lexiques HPSG et LTAG présentés dans les sections précédentes sont construits autour de 3 composantes conceptuelles : l'entrée lexicale, le mécanisme de typage et le mécanisme de relations entre entrées lexicales. Il est évident que le lexique HPSG est plus structuré que le lexique LTAG. Pour cette raison, nous utiliserons le formalisme HPSG comme formalisme pivot et nous convertissons les entrées lexicales du LTAG vers HPSG, en s'inspirant du travail de (Tateisi & al., 1998), avant de les transformer en LMF.

Nous partons des hypothèses suivantes :
– Le lexique est extentionnel. C'est à dire, il contient toutes les formes fléchies.
– Les règles lexicales sont prises en considération comme relations inter entrées et non pas comme des règles de génération de formes fléchies ou de formes dérivées. Nous nions le rôle productif aux règles lexicales.
– La hiérarchie de types contient seulement des informations de sous catégorisation. Nous supposons que les informations morphologiques sont développées sur les entrées lexicales [4].

La conversion est ordonnée : Nous débutons par la hiérarchie de types, puis les entrées lexicales et finalement les règles lexicales. nous utiliserons les règles suivantes afin d'aboutir à une description LMF.

Règle 1 Chaque type de la hiérarchie de types est projeté vers une instance du composant "SyntacticBehaviour" en LMF. Cette instance sera utilisée pour décrire le comportement syntaxique de plusieurs entrées lexicales.

[4] Normalement, les informations d'ordre morphologique sont embarqués dans les types quand c'est possible.

Règle 2 Chaque entrée lexicale HPSG est projetée vers une instance du composant "LexicalEntry" en LMF.

Règle 3 Chaque règle lexicale est projetée vers une instance du composant "LexicalEntry Relation" en LMF.

La projection d'un lexique HPSG vers LMF n'est pas une projection «un à un» dans la quelle chaque entrée lexicale HPSG se voit attribuer une entrée LMF correspondante. En fait, LMF est basé sur une vue sémasiologique des entrées lexicales et encode tous les sens d'un mot donné dans une seule entrée lexicale. De plus, LMF encode toutes les variantes morphologiques d'un mot donné dans la même entrée lexicale. De ce fait, le point de départ de la projection sera un ensemble d'entrées lexicales HPSG. La première étape à faire est de regrouper toutes les entrées lexicales HPSG possédant la même morphologie ou une morphologie déduite par l'application d'un paradigme flexionnel sur une racine morphologique. L'ensemble regroupé sera projeté en une seule entrée lexicale LMF fournissant une description des différents comportements syntaxiques, flexions et sens. La deuxième étape à faire est de transformer les informations morphosyntaxiques du trait de tête de chaque entrée lexicale en des propriétés morphologiques de l'instance du composant «Form». Un système de règles pour automatiser cette transformation est détaillé dans (Fehri & al., 2006).

La troisième étape est la projection du schéma valenciel. La structure argumentale dans la structure de traits indique le nombre et l'ordre des positions à instancier. Les traits de valence SUJ et COMPS fournissent les informations concernant les réalisations des positions et spécifiquement leurs catégories grammaticales (SN, SV,... indiquées explicitements dans le SAV) et leurs fonctions grammaticales (suj ou complément indiquées implicitement).

5. Discussion

La proposition de projection se base sur plusieurs hypothèses. En effet, le lexique HPSG est supposé être composé seulement d'entrées lexicales sans règles lexicales. Le rôle dérivationnel de ces règles est supposé, pour autant, existant. En plus, la projection de la description sémantique des entrées lexicales vers LMF a été ignorée bien que LMF dispose d'une extension sémantique. Finalement, un problème de perte de données lors de la projection d'un lexique HPSG en un lexique LMF a été rencontré (par exemple, la contrainte de 3eme personne au singulier pour le sujet a disparu dans la projection LMF).

La projection d'un lexique LTAG vers LMF en utilisant le formalisme HPSG comme pivot présente des risques de perte d'information. Ces risques doivent être étudiés pour pouvoir évaluer le lexique LMF résultant. Un chemin de conversion direct d'un lexique LTAG vers LMF pourra apporter plus de conformité du lexique résultant par rapport au lexique de départ. Mais il faudra identifier un système de règles de conversion de LTAG vers LMF.

6. Conclusion et Perspectives

Nous avons étudié la structuration et la représentation des lexiques LTAG et HPSG. Les similitudes et les correspondances conceptuelles entre les deux lexiques permettent de traduire les entrées lexicales LTAG vers HPSG. Nous avons par la suite donné une proposition pour la conversion du lexique HPSG vers LMF. Notre travail vise la création d'un lexique syntaxique dont la structure et la représentation seront universelles.

Comme perspectives, nous comptons prendre en considération le lexique du formalisme LFG (Lexical Functional Grammar) et créer une plateforme partagée pour la conversion et la fusion des lexiques syntaxiques existants.

References

ABEILLÉ A. (1991). Une grammaire lexicalisée d'arbres adjoints pour le français. application à l'analyse automatique. *Thèse de doctorat de linguistique, Université de Paris 7, 1991*.

AKROUT A. (2005). Modélisation d'un lexique flexionnel. application à l'arabe classique. Master's thesis, Université de Metz.

BAHOU Y. & AL. (2003). Vers une analyse syntaxique de la langue arabe basée sur une grammaire d'unification hpsg. *Troisième journées scientifiques des jeunes chercheurs en génie électrique et informatique, Mahdia, Tunisie*.

CANDITO M.-H. (1996). Organisation modulaire et paramétrable de grammaires électroniques lexicalisées. application au français et à l'italien. *Thèse de doctorat de linguistique, Université de Paris 7, 1996.*

DICHY J. (2003). Roots patterns vs. stems plus grammar-lexis specifications : on what basis should a multilingual lexical database centred on arabic be built ? *Proceedings of the MT-Summit IX workshop on Machine Translation for Semitic Languages. New-Orleans*.

ELLEUCH S. & AL. (2002). Adaptation des grammaires hpsg pour l'analyse de l'arabe. *Deuxième journées scientifiques des jeunes chercheurs en génie électrique et informatique, Hammamet, Tunisie*.

FEHRI H. & AL. (2006). un système de projection du hpsg arabisé vers la plate-forme lmf. *Journées d'études pour le traitement automatique de la langue arabe. Maroc*.

FRANCOPOULO G. & GEORGE M. (2005). *ISO/TC 37/SC 4 N130 Rev. 7 Language Ressource Management - Lexical Markup Framework (LMF)*. Rapport interne.

FRANCOPOULO . (2003). *Proposition de norme des lexiques pour le traitement automatique du langage*. Rapport interne, inria/loria-action syntaxe.

GAIFFE B. & AL. (2002). A new metagrammar compiler. *Proceedings of TAG+6, Venice*.

KROCH A. & JOSHI A. K. (1985). Linguistic relevance of tree adjoining grammars. *Technical report MS-CI-85-18*.

LOUKIL N. (2006). Une proposition de représentation normalisée des lexiques des grammaires d'unification. *RECITAL, Leuven.*

POLLARD C. & SAG I. (1994). *Head-Driven Phrase Structure Grammars*. Chigaco University Press.

TATEISI Y. & AL. (1998). Translating the xtag english grammar to hpsg. *In Proc. of TAG+4, pp. 172–175*.