Une proposition de représentation normalisée des lexiques HPSG

Noureddine Loukil

Laboratoire de Recherche en Informatique et Multimédia
noureddine.loukil@isimsf.rnu.tn

Résumé

L'interopérabilité entre les lexiques des grammaires d'unification passe par l'adoption d'une représentation normalisée de ces ressources. Dans ce papier, nous proposons l'utilisation de LMF pour établir la standardisation des ressources lexicales en HPSG. Nous présentons LMF d'une manière sommaire et nous détaillons son utilisation pour coder les entrées lexicales d'un lexique HPSG.

Mots-clés : lexique syntaxique, HPSG, Lexical Markup Framework, projection lexicale.

Abstract

The interoperability between unification grammar lexica can be established by adopting a normalized representation for those resources. In this paper, we propose the use of the Lexical Markup Framework to establish the standardization of HPSG lexical resources. We present an overview of LMF and we detail its use to code the lexical entries of an HPSG lexicon.

Keywords: syntactic lexicon, HPSG, Lexical Markup Framework, Lexical mapping.

1. Introduction

Les grammaires d'unification comme les grammaires lexicales fonctionnelles (Bresnan, 1982), les grammaires d'arbres adjoints (TAG) (Kroch et Joshi, 1985) et les grammaires syntagmatiques guidées par les têtes (Pollard et Sag, 1994) se basent sur une approche lexicalisée et essaient de coder l'essentiel de l'information linguistique dans leurs lexiques. Leurs entrées lexicales sont représentées sous forme de structures de traits (Carpenter, 1992), fournissant une spécification plus ou moins complète des caractéristiques linguistiques.

Les similitudes entre les grammaires d'unification au niveau du principe et de la représentation peuvent laisser entendre une exploitation commune des ressources lexicales de ces grammaires, voir une interopérabilité générale permettant à une ressource développée pour un formalisme donné d'être utilisée par les autres d'une manière transparente. Malheureusement, cela est loin d'être vrai. Les similitudes apparentes en surface cachent des différences de taille concernant généralement la méthode et la granularité de la description des ressources lexicales. L'élaboration d'une représentation standardisée pour les lexiques syntaxiques facilitera les développements futurs des lexiques pour les grammaires d'unification en établissant des convertisseurs automatiques permettant, en premier lieu, de projeter les ressources lexicales vers cette représentation standard, et en deuxième lieu, d'alimenter les nouveaux lexiques automatiquement à partir des ressources standardisées.

Dans cette optique, une proposition de normalisation est déjà en cours de discussion par le comité ISO TC 37/SC dans le cadre du projet LMF (Lexical Markup Framework) (Francopoulo, 2005), et ce pour élaborer le futur standard ISO 24613. LMF est le fruit de plusieurs travaux comme la fameuse initiative EAGLES (EAGLES, 96), Le projet MULTEXT (Calzolari, 1996) et le projet GENELEX (Antoni-Lay et Zaysser, 1994) qui ont essayé de proposer des modèles standard pour encoder l'information existante dans les différents types de lexiques TAL, voire même pour encoder les grammaires écrites pour les différents formalismes linguistiques. LMF propose non plus un modèle mais un méta modèle de représentation, permettant ainsi d'adopter les normes élaborées par les projets antérieurs. Le méta modèle abstrait proposé fournit un cadre standardisé pour la construction de nouveaux lexiques ou la restructuration de lexiques existants.

Dans ce papier, nous proposons l'utilisation de LMF pour établir la standardisation des ressources lexicales en HPSG, cela étant une étape d'un travail visant la généralisation pour toutes les grammaires d'unification. Ainsi, nous commençons par une étude sommaire de LMF, puis, nous détaillons son utilisation pour coder les entrées lexicales HPSG. Nous essayons, par la suite, d'établir un système de règles pour guider ce codage. Finalement, nous clôturons par une conclusion et des perspectives.

2. Aperçu sur LMF

LMF est basé sur une organisation sémasiologique des entrées lexicales. LMF est composé d'un méta modèle 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-modèle 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.

2.1. Le modèle de base

Le modèle de base de LMF, présenté à la figure 1, 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.

2.2. L'extension syntaxique de LMF

L'extension syntaxique de LMF est basée sur une étude approfondie de la sous-catégorisation dans les différents formalismes linguistiques. Étant un phénomène linguistique fondamental, la sous-catégorisation, a été bien étudiée dans l'initiative EAGLES et cela afin de fournir une description standardisé de ce phénomène indépendamment du formalisme. Les recommandations de EAGLES ont été adoptées dans LMF sous la forme d'une extension syntaxique au modèle de base. Dans ce qui suit, cette extension sera décrite et détaillée afin de dégager une méthodologie

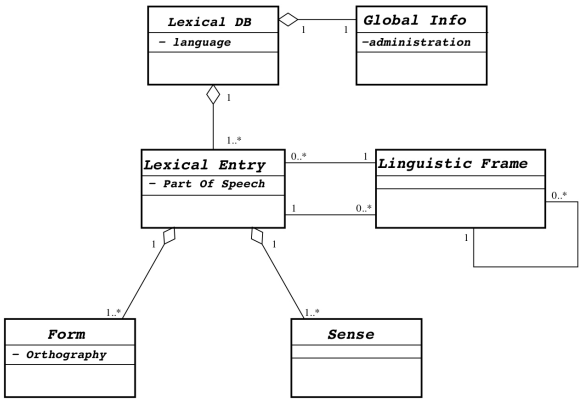

Figure 1. Le modèle de base de LMF

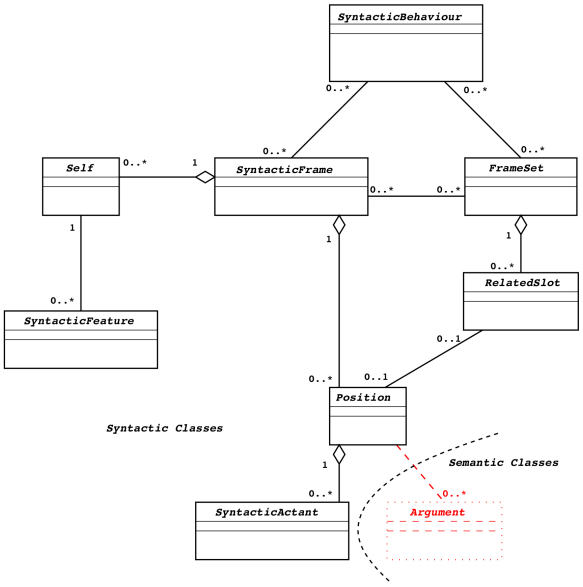

Figure 2. Extension syntaxique de LMF

de projection des informations dans une entrée lexicale dans un formalisme linguistique donné vers les éléments de l'extension. Le diagramme de l'extension syntaxique donnée à la figure 2 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. 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 conçu 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.

3. Le lexique HPSG

Dans cette section, nous présentons un exemple simple de deux entrées lexicales en HPSG. Dans la phrase de la figure 3, nous nous intéressons à la sous-catégorisation et à la structure sémantique du verbe arabe « *aabara* » "traverser". La figure 5 présente sa matrice attribut valeur en HPSG.

Le type valence comporte 3 traits appropriés : SUJ, COMPS et SPR. Le premier concerne le sujet. Sa valeur est une liste de valeurs de type SYNSEM. C'est également le cas du trait COMPS. Dans cet exemple, le verbe sous-catégorise un sujet et un objet direct. Ces valeurs se trouvent dans le trait S-ARG décrivant la structure argumentale. Ce dernier trait est en fait la concaténa-

aabara arrajulu azzukaka.
a traversé l'homme la rue.
« L'homme a traversé la rue. »

Figure 3. le verbe "aabara" (traverser)

aabara arrajulu.
a pleuré l'homme.
« l'homme a pleuré. »

Figure 4. le verbe "aabara" (pleurer)

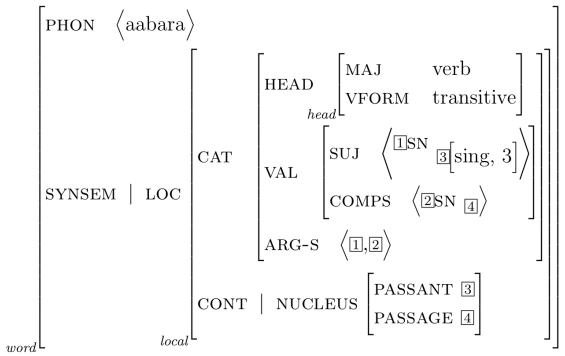

Figure 5. le verbe "aabara" (traverser)

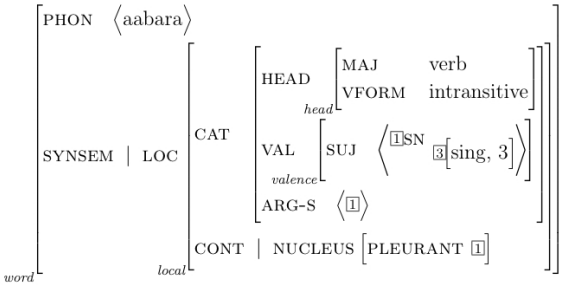

Figure 6. le verbe "aabara" (pleurer)

tion des traits de valence. De plus, le sujet porte une spécification sur son index indiquant que la catégorie nominale devra être à la 3ème personne du singulier. Les aspects sémantiques sont représentés dans le trait CONT qui comporte une liste de quantificateurs (vide puisqu'il s'agit d'une entrée lexicale). C'est pourquoi la structure sémantique sera représentée dans le trait NUCLEUS. Ce trait est composé de l'ensemble des arguments entrant dans la relation sémantique correspondante. Ainsi, dans notre exemple, la relation "passer" comporte deux arguments représentés par les traits "passant" et "passage" ayant pour valeur respectivement les traits d'index du sujet et du complément. Le même verbe possède une autre signification, phénomène courant dans plusieurs langues. La phrase de la figure 4 illustre cette situation. La différence au niveau de la sémantique implique généralement une modification dans les constructions syntaxiques permises par l'entrée. Dans cette situation, une nouvelle entrée lexicale est requise pour gérer la différence au niveau syntaxique, sémantique, voir même morphologique. Dans le SAV de la figure 6, le verbe sous-catégorise seulement un sujet. Sa valeur se trouve dans le trait S-ARG. Ce dernier trait est en fait la concaténation des traits de valence. De plus, le sujet porte une spécification sur son index indiquant que la catégorie nominale devra être à la 3ème personne du singulier. Pour la sémantique, la différence avec l'entrée précédente est dans le fait que le prédicat, « pleurer » dans ce cas, accepte un seul argument « PLEURANT » ayant pour valeur le trait d'index du sujet.

4. Projection des lexiques syntaxiques vers LMF

Plusieurs applications de LMF ont été réalisées pour la normalisation de bases de données lexicales flexionnelles (Romary et Francopoulo, 2004) et pour l'annotation des ressources linguistiques (Ide et de-la Clergerie, 2003). LMF définit les conditions qui permettent aux données, exprimées dans une ressource lexicale donnée, d'être projetées sur le framework LMF. Cependant, il ne définit pas la méthode avec laquelle cette projection doit être effectuée et ne fournit aucune garantie quant à sa correction. La suite de cette section essaie de proposer une méthode avec un exemple illustratif pour guider ce processus.

Dans les sections précédentes, nous avons présenté un mot unique -le verbe « *aabara* » possédant deux entrées lexicales HPSG distinctes correspondantes à deux sens et deux schémas

valenciels différents. Loin d'être une présentation exhaustive de la sous-catégorisation, cet exemple va nous servir pour prendre en considération le fait que 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 syntaxiques « SyntacticFeature » de l'instance « Self » appropriée. 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). La projection en LMF des entrées lexicales des figures 5 et 6 est donnée à la figure 7.

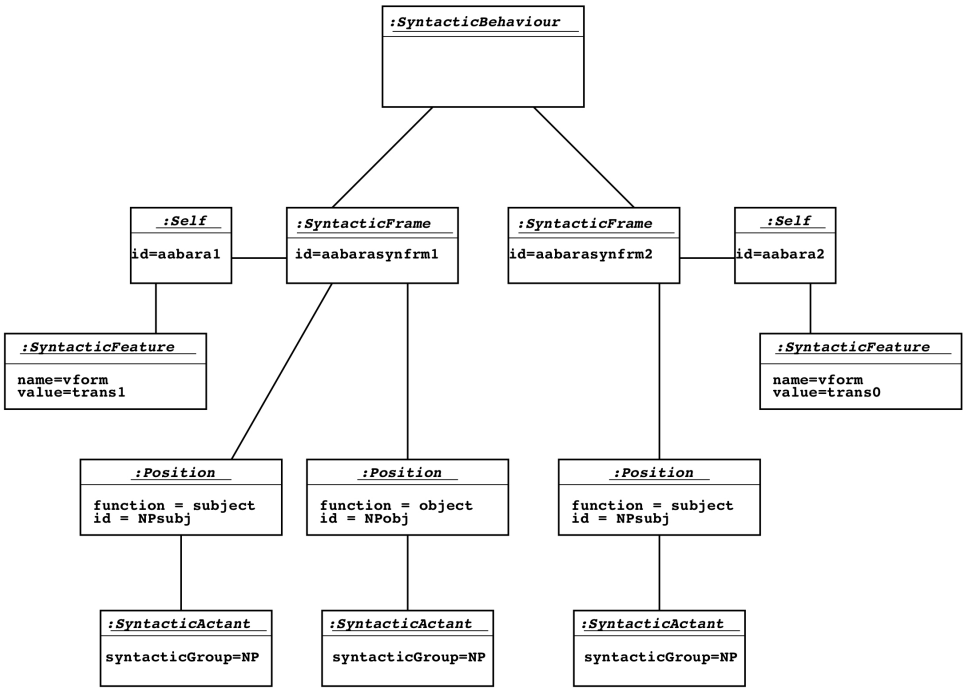

Figure 7. instance de l'extension syntaxique de LMF

La proposition de projection, étant simpliste, 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 3ème personne au singulier pour le sujet a disparu dans la projection LMF).

5. Conclusion et Perspectives

Dans ce papier, nous nous sommes limité aux ressources HPSG parce qu'elles offrent une représentation multi-niveaux pour chaque entrée lexicale. De plus, nous nous sommes abstenu d'étudier le codage des mots composés et des syntagmes pour réduire la complexité et introduire la projection. Notre but étant de bâtir des lexiques syntaxiques complets pour les grammaires d'unification à l'aide de LMF, ces limites seront abordées et résolues dans nos travaux futurs. Tout cela vise la création d'une structure modulaire qui permettra une vraie *interopérabilité* entre les lexiques syntaxiques de toutes les grammaires d'unification.

Références

ANTONI-LAY, MH. ; FRANCOPOULO G. et ZAYSSER L. (1994). « A generic model for reusable lexicons : The Genelex project ». In *Literary and Linguistic Computing*.

BRESNAN J. (1982). *The Mental Representation of Grammatical relations*. Cambridge, MA : MIT Press.

CALZOLARI, N. M. M. (1996). *Multext - Common Specifications and Notation for Lexicon Encoding*. Rapport interne.

CARPENTER B. (1992). « The Logic of Typed Feature Structures with Applications to Unification-based Grammars, Logic Programming and Constraint Resolution ». In *Cambridge Tracts in Theoretical Computer Science*.

EAGLES (96). *Reports of the Computational Lexicons Working Group*. Rapport interne.

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

IDE, N. ; ROMARY L. et DE-LA CLERGERIE E. (2003). « International Standard for a Linguistic Annotation Framework ». In *Proceedings of HLT-NAACL'03 Workshop on The Software Engineering and Architecture of Language Technology, Edmunton*.

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

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

ROMARY, L. ; SALMON-ALT S. et FRANCOPOULO G. (2004). « Standards Going concrete : from LMF to Morphalou ». In *Workshop on Electronic Dictionaries, COLING 2004, Geneva, Switzerland*.