

# **LEXICAL MARKUP FRAMEWORK : PRINCIPES FONDATEURS ET APPLICATION AUX LEXIQUES SYNTAXIQUES**

Susanne Salmon-Alt, Laurent Romary

Proposition de tutoriel

## 1. FONDEMENTS DU LEXICAL MARKUP FRAMEWORK

Le *Lexical Markup Framework* (LMF) [6] est une proposition de modélisation de données lexicales. Pour l’instant, il s’agit d’un item de travail (Working Draft 24613) de l’ISO TC 37/SC 4/WG 4, dédié aux ressources linguistiques et plus particulièrement aux bases de données lexicales. Cette proposition doit mener, à terme, à la définition d’une référence normalisée pour l’échange de données lexicales, qu’il s’agisse de données TAL ou de données dictionnairiques au sens classique.

Par rapport à d’autres initiatives de normalisation de lexiques, l’originalité de LMF provient de l’introduction d’une distinction explicite entre formats de codage et modèles de données sous-jacents. A titre illustratif, il est possible de comparer cette distinction avec celle que l’on préconise en génie logiciel entre la phase de conception (relevant de l’algorithmique) et la phase d’implémentation (relevant de la programmation). Beaucoup d’initiatives précédentes ont en effet tenté de proposer des standards directement sous forme de formats de codage, le cas prototypique étant la recherche de “la bonne DTD”. Or, l’inconvénient majeur d’une telle approche est un manque de flexibilité, puisque l’on impose aux utilisateurs à la fois le modèle de données sous-jacent (définition des propriétés des composantes et leur agencement) et le format de représentation (SGML/XML+DTD). Dans la lignée d’initiatives plus génériques, mises en oeuvre d’abord dans le domaine de l’annotation de corpus ([7], [8], [9], [5]), LMF repose sur l’hypothèse fondatrice que la standardisation de représentations linguistiques doit s’effectuer au niveau conceptuel plutôt qu’au niveau représentationnel.

LMF propose donc une modélisation conceptuelle des objets lexicaux sous forme d’un méta-modèle, associé à des catégories de données. Le méta-modèle reflète les propriétés structurelles des données : il s’agit d’un graphe orienté de structures de traits, caractérisant les objets linguistiques pertinents ainsi que les contraintes régissant leur agencement. Les catégories de données spécifient le contenu des données à représenter : ce sont des descripteurs linguistiques instanciant les traits et les valeurs. Leur gestion (description formelle, soumission, édition, documentation, standardisation, recherche et sélection) est indépendante de leur association effective avec un méta-modèle. Elle est externalisée et centralisée dans un registre de catégories de données, accessible en ligne (<http://www.syntax.loria.fr>).

L’association d’un ensemble de catégories de données aux noeuds d’un méta-modèle donne lieu à un modèle de données pleinement spécifié. Si LMF ne se préoccupe pas directement des formats d’instanciation (DTD, base de données relationnelle, tableaux ASCII, etc.), il propose néanmoins un format de représentation pivot en XML, parfaitement isomorphe au modèle.

## 2. ILLUSTRATION DES MÉCANISMES DE BASE

Partant du principe d'organisation lexicographique sémasiologique, les composantes fondamentales, sur lesquelles repose tout lexique conforme à LMF, sont l'*entrée lexicale*, dominant des *sens*, potentiellement hiérarchisés. A chacune des composantes sont attachés un certain nombre de descripteurs élémentaires (catégories de données) permettant de caractériser leur contenu. Mais la modularité du modèle LMF repose surtout sur la possibilité d'introduction d'extensions, venant se "greffer" sur les composantes du modèle noyau. Les extensions envisagées actuellement à l'ISO concernent le système morphologique (morphologie flexionnelle paradigmatique et extensionnelle), le système syntaxique (cadres de sous-catégorisation), le système sémantique (relations entre sens) et le système des traductions.

A partir de notre expérience acquise lors de la conception de *Morphalou*, un lexique flexionnel du Français libre et à large couverture [3], nous proposons pour ce tutoriel d'illustrer les mécanismes de base de LMF d'abord sur des données morphologiques. Sur la base d'un matériau linguistique relativement simple, cela permettra aux participants de se familiariser avec

## 3. DISCUSSION DE L'EXTENSION SYNTAXIQUE

Suivant l'approche sémasiologique, LMF part du postulat de la primauté du sens. Par conséquent, la première hypothèse pour la représentation de structures syntaxiques est leur subordination au(x) sens. Ainsi, un vocable polysème pourra être associé à au moins autant de structures syntaxiques que de sens. En terme de modélisation, cela signifie que l'extension syntaxique de LMF se greffe sur une composante *sens* existante.

Dans sa conception actuelle, l'extension syntaxique couvre essentiellement la description des structures argumentales, ou cadres de sous-catégorisation, pour des entrées lexicales à sens prédicatif, notamment les verbes. En attendant une stabilisation autour des composantes à retenir définitivement, l'extension a été conçue dans une approche ascendante, partant de la modélisation des "observables" effectifs, et rendant ainsi possible une abstraction par étapes successives.

**Position Syntaxique.** Les "observables" directs au niveau syntaxique sont les occupants des positions argumentales offerte par une lexie prédicative. Pour décrire celles-ci, LMF prévoit donc d'abord une composante *syntacticPosition*, régissant les catégories de données suivantes : *syntacticFunction* pour la fonction grammaticale (sujet, objet, etc.), *syntacticConstituent* pour la forme du constituant (groupe nominal, groupe prépositionnel, subordonnée etc.), *introducer* pour l'élément introducteur de certains types de constituants (préposition ou conjonction de subordination). Ce tryptique de descripteurs basiques correspond à un noyau stable d'attributs que l'on retrouve dans les initiatives de modélisation précédentes majeures (Genelex<sup>1</sup>, EAGLES<sup>2</sup>, Parole<sup>3</sup>).

D'autres descripteurs méritent un examen plus attentif avant d'être intégrés dans LMF : la question de l'introduction d'une catégorie de données spécifique *syntacticCase* (s'appliquant uniquement au noms) est encore ouverte et dépendra de l'articulation entre celle-ci et *syntacticFunction*. La sémantique des attributs *definition*, *comment* et *example* (EAGLES) doit non seulement être spécifiée ("Qu'est-ce

---

<sup>1</sup>[4]

<sup>2</sup>[1]

<sup>3</sup>[2]

que la définition d'une position syntaxique ?"), mais aussi être ré-examinée à la lumière d'une éventuelle articulation d'un lexique avec des données observées en corpus : il est par exemple imaginable de vouloir encoder non seulement un exemple (cas typique des dictionnaires ou lexiques sources, comme le TLFi ou les tables du LADL), mais aussi de vouloir faire le lien avec un ensemble de réalisations dans un corpus de référence (cas de l'acquisition de cadres de sous-catégorisation en corpus). Enfin, l'articulation avec des informations provenant d'autres niveaux de description, ou d'un cadre de sous-catégorisation, reste sujet à discussion. Dans un premier temps, nous n'avons pas retenu les *restrictions sélectionnelles* comme descripteur d'une position syntaxique, puisque ces renseignements (le plus souvent *+/-animé* ou *+/-humain*) caractérisent un argument sémantique plutôt que syntaxique. Ceci étant, et à l'instar d'autres propositions (EAGLES), nous avons maintenu la possibilité d'apparier une position syntaxique à un argument sémantique par un mécanisme de pointage. Il est toutefois à noter que ce lien n'est ni systématique, ni bi-univoque. Un dernier point de discussion est le *statut* (facultatif vs. obligatoire) des positions syntaxiques décrites. La pertinence d'un tel descripteur est en effet fortement corrélée au degré d'abstraction (ou de factorisation) que l'on accorde à la description d'un ensemble de position (i.e. d'une construction syntaxique).

**Construction Syntaxique.** Le premier niveau d'abstraction au-dessus des "observables" est la *syntactic Construction*, correspondant à une cadre de sous-catégorisation. Nous définissons une construction syntaxique comme un ensemble de positions syntaxiques réalisables simultanément pour une même lexie verbale<sup>4</sup>. L'introduction d'une telle composante se justifie par le fait qu'elle fournit un point d'ancrage pour certaines informations syntaxiques relatives à la lexie verbale elle-même ou à l'ensemble des positions regroupées. Parmi ces informations, on peut mentionner l'auxiliaire verbal du participe passé (*être/avoir*) ou des contraintes sur la réalisation de certaines positions en fonction de la réalisation d'autres positions (cf. Genelex).

Un des points à discussion est la portée d'une constructions syntaxique ("Considère-t-on un ensemble maximal de positions ?"). Une réponse positive à cette interrogation - réponse adoptée par une majorité de travaux précédents - entraînerait l'introduction de marqueurs d'optionnalité pour certaines positions réalisées facultativement, comme dans *X noye Y (dans Z)*. Parmi les autres questions ouvertes, la question du marquage explicite de l'ordre des positions syntaxiques peut trouver une réponse (négative), en suivant l'argumentation en faveur de l'adoption d'un ordre conventionnel (cf. Genelex). Enfin, comme pour les positions syntaxiques, le souci de l'articulation de la description des constructions avec des données observées en corpus peut être pris en compte par des catégories de données spécifiques.

**Ensemble de constructions.** Au delà de la construction syntaxique, plusieurs initiatives introduisent une composante permettant de regrouper un certain nombre de constructions syntaxiques. Un tel ensemble (*FrameSet* dans EAGLES/PAROLE) correspond à une abstraction supplémentaire, dans la mesure où il factorise la description de certaines constructions syntaxiques comme résultant d'alternances syntaxiques régulières (passivation, ergativisation etc.). Le rapport PAROLE introduit cette composante comme un moyen de représenter des informations relevant de la

---

<sup>4</sup>Notre définition est plus précise que celle adoptée dans Genelex. Elle exclut par exemple la possibilité de traiter *Pierre répond à la question.* et *Pierre répond que c'est vrai.* comme relevant d'un même cadre, ce qui nous paraît discutable (cf. Antoni-Lay et al. (1994), p. 14)

“syntaxe profonde” au sens chomskyien. Pour un verbe tel que *cuire*, il serait ainsi possible de ne décrire qu’une seule cadre pour des réalisations syntaxiques différentes (*Jean cuit le poulet.* et *Le poulet cuit.*)

#### 4. PROPOSITION DE PLAN POUR LE TUTORIEL (SUR LA BASE DE 1H30)

##### 4.1. Normalisation de ressources lexicales : 10’.

- Objectifs et contexte (ISO TC 37/SC 4)
- Principes fondateurs (séparation “concepts” et “formats”)

##### 4.2. LMF : mécanismes de base : 30’.

- Notions essentielles (présentation du méta-modèle, catégorie de données, extension lexicale)
- Gestion des catégories de données (démonstration en ligne du registre des catégories de données)
- Choix de modélisation pour des aspects linguistiques non consensuels (travail pratique sur les féminisations)
- Instanciation concrète des représentations lexicales (présentation du XML pivot et différentes représentations utilisateur)

##### 4.3. L’extension syntaxique : 60’.

- Initiatives “sources” (présentation Genelex, EAGLES, PAROLE, ISLE)
- Transfert et intégration dans LMF (présentation des composantes essentielles)
- Discussion des questions ouvertes (à partir de la mise en commun d’un travail pratique consistant à transférer du matériel existant vers LMF : extraits des tables du LADL, du TLFi, échantillons acquis en corpus, lexiques d’analyseurs syntaxiques, matériel apporté par les participants)

#### REFERENCES

- [1] Eagles - recommendations on subcategorization. <http://www.ilc.cnr.it/EAGLES96/segsasg1/segsasg1.html>, 1996.
- [2] Parole - report on syntactic layer. [http://www.ub.es/gilcub/SIMPLE/reports/parole/parole\\_syn/parosyn\\_2.html](http://www.ub.es/gilcub/SIMPLE/reports/parole/parole_syn/parosyn_2.html), 1998.
- [3] Romary L., S. Salmon-Alt, and G. Francopoulo. Standards going concrete : from lmf to morphalou. In *Workshop on Electronic Dictionaries, Coling 2004*, Geneva, Switzerland, 2004.
- [4] Antony-Lay M.-H., G. Francopoulo, and L. Zaysser. A generic model for reusable lexicons: the genelex project. *Literary and Linguistic Computing*, 9(1):47–54, 1994.
- [5] Ide N. and L. Romary. A registry of standard data categories for linguistic annotation. In *Proceedings of LREC*, Lisbon, 2004.
- [6] ISO TC 37/SC 4 N130 Rev.3. Language resource management - lexical markup framework. In *Working Draft of ISO WD 24613:2004*, 2004.
- [7] Bird S. and M. Liberman. A formal framework for linguistic annotation. *Speech Communication*, 33(1-2):23–60, 2001.
- [8] Davies S. and M. Poesio. Coreference. In *MATE Dialogue Annotation Guidelines, Deliverable D2.1*, 2000.
- [9] Salmon-Alt Susanne. Du corpus à la théorie : l’annotation (co-)référentielle. *Traitement Automatique des Langues*, 42(2), 2001.