V[1]Ω a=able
ou
Normaliser des lexiques syntaxiques est délectable[1]

Susanne Salmon-Alt

ATILF-CNRS, Nancy
salt@atilf.fr

Résumé

Partant des lexiques TAL syntaxiques existants, cet article propose une représentation lexicale unifiée et normalisée, préalable et nécessaire à toute exploitation des lexiques syntaxiques hors de leur propre contexte de conception. Ce travail s'inscrit dans un cadre de modélisation privilégié – le *Lexical Markup Framework* – qui a été conçu dès le départ comme un modèle lexicographique intégrant les différents niveaux de description. Ce modèle permet d'articuler des descriptions extensionnelles et intensionnelles et fait référence à un jeu de descripteurs normalisés, garantissant la rigueur de la description des faits linguistiques et assurant, à terme, la compatibilité avec des formats de données utilisés pour l'annotation de corpus.

Mots-clés : lexique, TAL, syntaxe, lexique-grammaire, sous-catégorisation, standardisation.

Abstract

Based on existing lexical resources for NLP, in particular inflected form and subcategorization lexica, this paper proposes a unified and normalized representation, required for any further use of the data out of their original context. As a starting point for our model, we chose the Lexical Markup Framework, for three reasons. Firstly, it covers various layers of linguistic description including morphology, syntax and semantics. Secondly, it allows for combining extensional (*i.e.* lists of forms or constructions) and intensional (*i.e.* reference to paradigms) lexical descriptions. Thirdly, it makes use of externally defined data categories, ensuring linguistic soundness and, ultimately, compatibility with standardized corpus annotation formats.

Keywords: lexicon, NLP, syntax, lexicon-grammar, normalization, subcategorization.

1. Normalisation de structures lexicales : motivations et cadre

Si la communauté francophone du TAL s'est dotée, notamment durant ces dernières années, de lexiques de formes fléchies libres et suffisamment importants pour à la fois encourager un travail de normalisation et y servir de support, la situation est différente pour les lexiques syntaxiques. Les mises en ligne d'un certain nombre de données[2] – par exemple LEFFF2 (Sagot *et al.,* 2006, Proton (Van den Eynde et Mertens, 2003) ou le Lexique-Grammaire – peuvent être considérées comme un pas dans la bonne direction. Toutefois, au vu de la divergence théorique et de la complexité de ces données syntaxiques, la question d'un modèle de données commun se pose dès lors que l'on souhaite comparer, fusionner, interfacer et documenter ces données.

[1] La formule est empruntée aux en-têtes de la table 4 du Lexique-Grammaire (Gross, 1975 : 252). Son application à une entrée lexicale particulière, *e.g. délecter*, permet de générer des phrases telle que celle du titre.

[2] Les URL sont indiquées dans les références bibliographiques.

Dans la continuité de la modélisation lexicale pour des lexiques flexionnels (Romary *et al.*, 2004), il nous a semblé intéressant d'explorer la modélisation des lexiques syntaxiques dans un cadre unifié posé par le *Lexical Markup Framework* (LMF, Monte et Francopoulo, 2005). Il s'agit d'une proposition dont l'ambition est de devenir, à terme, une référence normalisée pour la représentation de structures lexicales dans une perspective sémasiologique (future norme ISO-24613). Outre sa modularité et sa généricité vis-à-vis de modèles ou de théories lexicologiques particuliers, LMF se caractérise essentiellement par l'abstraction par rapport au langage de représentation effectif (SGML/XML, DTD propriétaire ou TEI, base de données relationnelle, etc.).

Dans une perspective de généricité, la modélisation opère en effet au niveau conceptuel : elle vise à inventorier les composantes essentielles d'un modèle lexicologique ainsi qu'à décrire un jeu de contraintes minimales régissant leur agencement. L'ensemble de composantes peut être considéré comme un méta-modèle au sens d'UML, c'est-à-dire un modèle qui décrit un ensemble de modèles par l'énonciation explicite des principes et règles nécessaires à la conception d'un modèle dans un domaine particulier. Par analogie avec la notion de modèle en sémantique formelle, on peut considérer un méta-modèle comme une fonction qui assigne aux termes d'un langage formel non pas des dénotations, mais des types d'entités sémantiques.

La modularité de LMF est assurée par les mécanismes d'adjonction, aux composantes du méta-modèle, de descripteurs élémentaires (catégories de données) et/ou de composantes optionnelles (extensions lexicales). Les catégories de données correspondent à des notions linguistiques descriptives élémentaires (par exemple, *catégorie grammaticale*, *genre*, *pluriel*). Elles sont inventoriées et définies, dans un cadre formel emprunté à la description terminologique (Ide et Romary, 2004), dans un registre de catégories de données (DCR), consultable et éditable en ligne. Une catégorie de données n'est donc pas associée de façon normative à une composante du méta-modèle, ni même limitée à un usage dans le cadre de LMF : un des avantages majeurs de cette gestion réside dans la maintenance d'un seul référentiel de descripteurs linguistiques normalisés, utilisables à la fois pour la description de lexiques et l'annotation de corpus.

Sur la base de ces principes, nous avons spécifié et implémenté une structure de données pour la représentation de lexiques syntaxiques (section 2) qui sert actuellement de cadre de prototypage pour la représentation normalisée d'un lexique syntaxique dérivé des tables du Lexique-Grammaire (section 3). La mise en perspective (section 4) aborde la question du déploiement de LMF pour une base commune de lexiques syntaxiques du français.

2. Modéliser des lexiques de sous-catégorisation

2.1. L'extension syntaxique LMF

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 de l'entrée lexicale. Dans sa conception actuelle, l'extension syntaxique couvre essentiellement la description des structures argumentales. En attendant une stabilisation autour des composantes à retenir définitivement, nous proposons une approche de modélisation ascendante (Figure 1), partant des « observables » effectifs, et rendant ainsi possible une abstraction par étapes successives.

Les premiers « observables » syntaxiques sont les occupants des positions argumentales d'une lexie prédicative. Pour décrire celles-ci, LMF prévoit une composante /*syntacticArgument*/,

que nous proposons de définir comme une « position structurelle nominale reliée à un même prédicat, caractérisée selon un nombre limité de types de comportements syntaxiques » (Creissels, 1995). Cette définition appelle trois remarques. Premièrement, la notion de « position structurelle nominale » désigne toute position pouvant potentiellement être occupée par un groupe nominal. À cette condition, elle inclut les complétives subordonnées ou infinitives. La forme de réalisation de l'argument – sa catégorie syntaxique ainsi que des introducteurs éventuels – est précisément spécifiée par des catégories de données appropriées, *i.e.* /*syntacticCategory*/ et /*syntacticIntroducer*/. Deuxièmement, la nature de la relation au prédicat est volontairement laissée sous-spécifiée. En particulier, LMF ne prend pas position vis-à-vis de la dichotomie théorique argument–circonstanciel qu'on sait contestée d'un point de vue typologique (Creissels, 1995) et psycholinguistique (Koenig *et al.*, 2003). Dans cette perspective, la définition du type de relation à considérer relève de décisions éditoriales et doit être adaptable au cas par cas. Troisièmement, les types de comportements syntaxiques sont communément résumés par les fonctions syntaxiques (sujet, objet etc.), dont les valeurs doivent faire l'objet de définitions claires dans des implémentations précises. En particulier, nous ne préconisons pas l'introduction d'un marqueur du cas grammatical (nominatif, datif etc.). Il s'agit d'un trait morphologique à décrire au niveau flexionnel, qui ne doit pas être confondu avec la fonction syntaxique du constituant en question : Dans les langues à cas, les cas grammaticaux ne correspondant pas systématiquement à des relations prédicat-argument (cf. le génitif allemand). En contrepartie, il est prévu deux catégories de données relationnelles pour encoder des liens d'identité entre arguments syntaxiques (cas des « contrôleurs » ou de l'appariement des arguments en diathèse) et des liens entre arguments syntaxiques et sémantiques.

L'ensemble des composantes représentant les arguments (ainsi qu'une composante optionnelle pour le prédicat, utile pour la description récursive des arguments phrastiques) forme un cadre syntaxique, *i.e.* un « sous-ensemble de la combinatoire de positions syntaxiques » (Antoni-Lay *et al.*, 1994). En s'appuyant sur les arguments syntaxiques plutôt que sur les fonctions syntaxiques, cette définition esquive volontairement la question de l'identification unique d'un cadre syntaxique : le fait de faire relever *Pierre répond*, *Pierre répond à la question* et *Pierre répond que c'est vrai* d'un seul, de deux ou de trois cadres doit rester paramétrable par les théories syntaxiques sous-jacentes (*e.g.* la prise en compte d'opérations de transformation ou non) et la granularité des fonctions et des constituants. Ce qui importe est la capacité du modèle à sous-spécifier certaines informations et à encoder des constructions de façon extensionnelle ou intensionnelle. Or, la représentation par intension s'intègre facilement dans LMF, comme nous l'avons déjà mis en œuvre au niveau flexionnel (Salmon-Alt *et al.*, 2005) : de la même façon qu'un composant abstrait /*form*/ peut s'instancier en tant que forme fléchie ou lemmatisée, nous considérons /*syntacticFrame*/ comme un composant abstrait, s'instanciant au choix en tant que construction effective (« surfacique ») ou abstraite (« profonde »). La description d'une construction abstraite, associée à un paradigme de transformations syntaxiques capables de « générer » la liste des constructions effectives, permet de faire l'économie d'une représentation extensionnelle, à condition de prévoir l'association, aux règles, de scores de productivité et de contraintes sur leur ordonnancement, comme cela est par exemple possible pour les grammaires d'unification. La majeure difficulté reste d'ordre théorique et concerne le choix des constructions profondes et des règles de transformation.

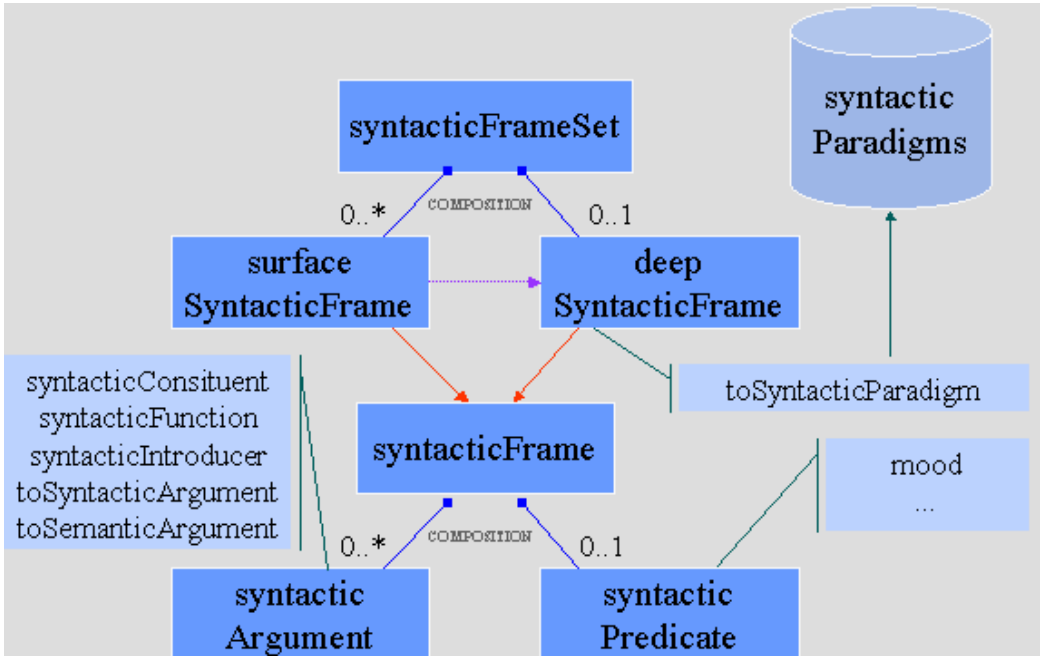

Figure 1. LMF – Spécification pour lexiques de sous-catégorisation

2.2. Un exemple d'instanciation

Le lexique Proton est une base de données de valences verbales pour le français, conçu selon l'approche pronominale (van den Eynde et Mertens, 2003). Comme le montre la Figure 2, elle présente, pour chaque schéma valenciel d'un verbe, en plus des exemples et traductions, une description des positions argumentales [arg] associées à des fonctions syntaxiques [p0, p1 etc.] ainsi que des opérations de transformation [reform]. L'originalité de l'approche consiste en la caractérisation des arguments en fonction des possibilités de pronominalisation, et à en déduire leur catégorie grammaticale et des traits sémantiques.

```
entry( désirer, 28350,
 example('r : mais moi, je désire tout ...'),
 translation(du, 'verlangen (naar), wensen, willen hebben, begeren'),
 [
  arg( p0, obl,
       paradigm( [je, nous, on, qui, elle, il, ils, 'celui-ci', 'ceux-ci'] ),
       features([human: +, cat:nominal]) ),
  arg( p1, obl,
       paradigm( [te, vous, qui, ceci, 'le(qpsubj)', 'le(inf)', 'ça(qpsubj)', 'ça(inf)', la, le, les, 'en Q', que, 'celui-ci', 'ceux-
                  ci', ça, 'l\'un l\'autre', 'se réc.'] ),
       features([human: +/-, cat:nomiverb, mood:subj]) )
 ],
 reform(['passif être', 'se passif'])
).
```

Figure 2. Entrée désirer *dans Proton (structure de données Prolog)*

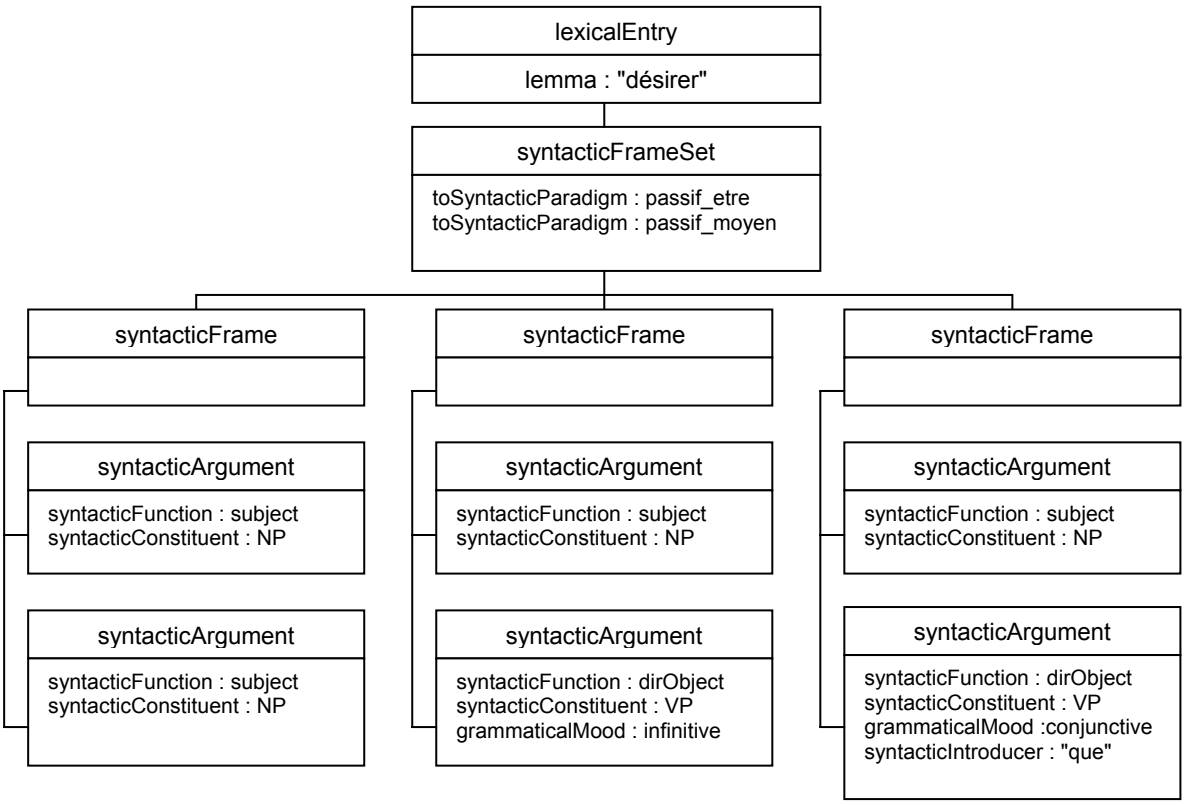

Figure 3. Entrée désirer *de Proton en LMF*

La conversion de cette entrée vers un format LMF (Figure 3) part de la création d'une entrée lexicale pour le lemme. L'ensemble des cadres – instanciés ici simplement comme /*syntacticFrame*/ sans spécifier leur nature canonique ou surfacique – sont regroupés sous un /*syntacticFrameSet*/. La question la plus importante est le découpage en cadres individuels : elle pose le problème de l'identification unique d'un argument, par exemple lorsqu'un argument peut se réaliser par deux catégories syntaxiques différentes (groupe nominal et verbal), ou lorsque certaines constructions imposent des contraintes supplémentaires sur la réalisation d'un groupe nominal (construction réciproque). En l'absence de décisions éditoriales spécifiques, la perspective extensionnelle préconise la création de cadres distincts, afin d'éviter que la combinatoire d'alternatives sur plusieurs arguments en parallèle mène à des cadres incorrects. Au delà de l'instanciation des traits élémentaires pour caractériser les arguments – relativement facilement convertibles à partir du lexique source –, il sera intéressant de sauvegarder les informations linguistiques spécifiques qu'offre Proton, *i.e.* le profil des pronominalisations et les propriétés sémantiques des arguments. Enfin, pour l'attachement des opérations de transformations (passif être, passif moyen), il faudra choisir une composante d'ancrage : en l'absence de critères permettant d'identifier une construction particulière en tant que construction canonique, il semble raisonnable de les adjoindre à l'ensemble des cadres, *i.e.* au /*syntacticFrameSet*/.

L'application de LMF à un lexique particulier demande donc essentiellement deux choses : d'une part, à apparier les descripteurs propriétaires avec les catégories de données normalisées (*p0 => subject, cat => syntacticConstituent* etc.), et d'autre part, à effectuer des choix linguistiques concernant le degré de granularité et de factorisation de la description des entrée lexicales, constructions et arguments syntaxiques. Ce dernier aspect fait l'objet d'une discussion plus détaillée dans la section suivante.

3. Étude de cas : le Lexique-Grammaire

Gardent *et al.* (2005) proposent un lexique de sous-catégorisation pour le français, généré automatiquement à partir de certaines tables des constructions complétives du LADL (Gross, 1975). Le travail part du constat que l'interfaçage du lexique-grammaire dans sa forme originale avec des applications TAL est « hampered both by its non standard encoding and by a structure that is partly implicit and partly underspecified ». Nous examinerons donc certaines propriétés de ce nouveau lexique[3] sous ces deux aspects (caractère standardisé et explicite), en suggérant que le repositionnement d'un tel travail dans le cadre de LMF pourrait ouvrir des voies de réflexion menant à des résultats plus proches des objectifs visés.

3.1. La macro-structure et la notion de lemme

Traditionnellement, un dictionnaire sémasiologique organise les entrées selon les lemmes. Le lemme peut être défini comme une forme canonique et abstraite sur les occurrences en corpus qui sont les formes fléchies, à laquelle on associe des informations sémantiques.

Au vu de cette définition, *SynLex* n'identifie pas clairement le statut linguistique de ses entrées : le point d'entrée n'est pas la forme canonique, puisqu'une même forme peut donner lieu à plusieurs entrées distinctes (cf. 3 entrées pour *accorder*). Ce traitement pourrait être rapproché du traitement classique de l'homonymie, mais pose problème à plusieurs égard. D'abord, en créant autant d'entrées que de sens, on systématise l'homonymie au détriment de la polysémie, ce qui peut être un choix délibéré, mais mérite d'être explicité. Par ailleurs, *LexSynt* se différencie d'un traitement sémasiologique classique par le fait qu'une même entrée peut traiter de plusieurs formes canoniques, par exemple en cas de dérivation adjectivale (cf. l'entrée *ennuyer* : *ennuyer, ennuyeux, ennuyant*). Ce choix limite fortement l'interopérabilité avec d'autres lexiques – considérant un lemme comme le représentant d'un paradigme flexionnel unique – et surtout, ne se justifie pas tant que la question de la sous-catégorisation de ces dérivés (cf. le transfert des cadres au cours d'une opération dérivationnelle) n'est pas abordée. L'absence d'une implémentation claire de la notion de lemme se reflète également dans le traitement des verbes pronominaux. L'entrée des ces verbes a été réduite à une forme simple, la particule étant encodée dans un trait *particule* dupliqué pour chaque cadre. Or, la relégation de la description d'un fait lexicalisé au niveau des cadres syntaxiques, outre la redondance, introduit surtout une ambiguïté absente du lexique original, à savoir la non-distinction formelle entre verbes intrinsèquement pronominaux et constructions réflexives.

Par ailleurs, au vu du traitement d'autres entrées complexes (verbe + adverbe, particule de négation ou *en*), il convient de s'interroger sur le statut linguistique accordé à la classe des *particules* (comprenant *en, ne, n', n'en, se, s'* ou *s'en*) et à certaines entrées hétéroclites, telles que *trouver Advm* ou *plaindre pas* (associé à une particule *ne se*). Cette question est éminemment liée à celle de l'identité du lemme, puisqu'elle concerne la description des mêmes objets lexicaux et se pose précisément aux confins de la définition traditionnelle du lemme : elle touche à la fois à la nature grammaticale (catégorie et propriétés flexionnelles) et la (dé-)composition formelle de lexies complexes, figées ou semi-figées. En l'absence de solutions standardisées toutes prêtes, différents traitements moins hétérogènes et fondés sur

[3] SynLex : en l'absence de métadonnées appropriées (versionnage et mises à jour), il n'est pas possible de référencer une version stable du lexique diffusé. Les données sur lesquelles nous fondons nos observations sont celles du 5 et 6 décembre 2005.

des arguments linguistiques sont envisageables et ont été discutés par exemple dans Geyken et Boyd-Graber (2003).

3.2. La description des constructions syntaxiques

SynLex décrit les constructions syntaxiques dérivées du Lexique-Grammaire par des structures de traits : une construction correspond à un ensemble d'arguments, caractérisés par des traits pour la fonction syntaxique, la réalisation syntaxique, les introducteurs, les contrôleurs et des contraintes sémantiques et/ou morphosyntaxiques. Si cette structure est proche de ce que propose LMF, certains problèmes résident dans l'absence de définition des étiquettes et dans l'inadéquation de leur combinatoire : à titre d'exemple, les arguments en fonction de *DeObjet* possèdent de façon redondante un introducteur prépositionnel *de,* état de fait injustifié en l'absence d'une définition en intension des *DeObjet* qui fasse apparaître des propriétés (*e.g.* de pronominalisation) particulières.

Par ailleurs, les arguments de la forme *de ce Que P* (*Max se félicite de ce qu'Ida ne vient pas*.) sont caractérisés comme constituants phrastiques, introduits simultanément par une préposition *de* et un complémenteur *ce que* (*féliciter–8*, cf. annexe). Or, la motivation linguistique de cette analyse (ainsi que celle des autres complémenteurs) n'est pas immédiate : elle masque la nature nominale du pronom démonstratif en tête du constituant, elle crée une unité linguistique fantôme *ce que* appartenant à une classe hétérogène de *complémenteurs*, et elle s'oppose à des arguments linguistiques en faveur d'une distribution complémentaire des prépositions et conjonctions de subordination (Creissels, 1995 ; Tesnière, 1959).

Un point particulièrement délicat est la gestion simultanée des représentations extensionnelle et intensionnelle. Les cartouches des tables du Lexique-Grammaire décrivent en effet des constructions de façon explicite (*N1 est Vpp de ce Qu P*) ou implicite, *i.e.* par des règles de transformation ([*passif par*]). Opérer en mode intensionnel signifie alors d'identifier les constructions de base auxquelles les transformations s'appliquent. Pour le Lexique-Grammaire, on peut supposer raisonnablement qu'elles s'appliquent aux constructions définitoires des tables (Gross, 1975 : 232). Or, dans *SynLex*, les constructions définitoires ne sont pas formellement identifiées, ce qui soulève certaines difficultés. À titre d'exemple, mentionnons des applications de passivation à des constructions autres que définitoires qui mènent à des constructions douteuses, telles que pour *graver* de la table 10. Cette table comporte des verbes à deux compléments, le premier étant direct et potentiellement phrastique, le second étant indirect et différent de *à + GN humain.* La construction définitoire de cette table ainsi qu'un exemple qui l'instancie pour *graver* sont donnés en (1). La table prévoit, pour ce verbe, une transformation passive, illustrée en (2). Par ailleurs, le sujet est susceptible d'être réalisé sous forme de complétive, comme en (3). Toutefois, appliquer la passivation à la construction résultante de (3) aboutit à (4) qui semble douteux. C'est pourtant ce que prévoit *SynLex* par l'entrée (5) pour *graver*.

1. N_0 V Qu Prép N_2 *Max a gravé dans la mémoire que Léa ne reviendrait pas.*
2. [passif par] *Que Léa ne reviendrait pas a été gravé par Max dans la mémoire.*
3. N0 =: V2c Ω *De lire ce cette lettre a gravé dans sa mémoire que Léa ne reviendrait pas.*
4. [3] + [2] *? Que Léa ne reviendrait pas a été gravé par lire cette lettre dans sa mémoire.*

5. p[arg=a0, fonc=sujet, **mode=inf, controleur=a2c**|a1] v n[arg=a2, fonc=locatif, type_sem=locatif_abstrait, prep=dans] n[arg=a1, fonc=objet, type_sem=non-humain] **passif_par**

La co-indexation des arguments dans des constructions diathétiques est également problématique : le cadre du passif, par exemple, utilise le trait *arg* non plus comme identifiant, mais comme renvoi à un argument d'une hypothétique construction canonique qui n'est pas identifiable. Cette double sémantique de *arg* est non seulement implicite et difficilement opératoire, mais pose plus largement la question de la co-indexation comme opération sémantique, présupposant idéalement une représentation sémantique associée à des constructions canoniques.

4. Lexiques syntaxiques et LMF : perspectives de déploiement

Repositionner ces questions dans le cadre d'un modèle lexicographique holistique et normalisé n'apporte pas toujours des réponses faciles et opératoires, mais revient à se donner une grille de lecture qui guide l'analyse de données initialement complexes et aide à repérer des incohérences. Si beaucoup de données problématiques discutées *supra* figurent telles quelles dans le Lexique-Grammaire, nous considérons effectivement que la plus-value d'un lexique TAL explicite et standardisé devrait précisément résider dans le questionnement des données, dans l'implémentation et la documentation des choix d'explicitation, ainsi que dans la diffusion des résultats sous un format lexicographique adéquat et pérenne.

D'un point de vue de la macro-structure, l'hypothèse sous-jacente au Lexique-Grammaire (une ligne = un sens) s'accorde parfaitement avec le principe sémasiologique de LMF : les entrées sont des associations d'une forme canonique à *n* sens, le nombre de sens correspondant au nombre de lignes du lexique-grammaire. Si l'on ne souhaite pas introduire une composante /*sens*/ explicite, l'utilisation itérative de /*syntacticFrameSet*/ à l'intérieur d'une entrée lexicale semble tout à fait appropriée. Par ailleurs, une orientation clairement sémasiologique suggérerait de traiter les dérivés adjectivaux dans une entrée à part, à moins d'élargir considérablement (et abusivement) la notion de lemme. Partant d'une définition plus claire du lemme, une solution simple et conforme à LMF pour le traitement des entrées complexes consisterait à les encoder par des formes composées (*se féliciter*, *trouver bien*, *trouver mal*, *ne pas se plaindre*) qui, au besoin, pourraient être décomposées à leur tour.

La nécessité d'identifier une construction canonique – correspondant au /*deepSyntacticFrame*/ dans LMF – conduit naturellement à attribuer ce rôle à la construction définitoire du Lexique-Grammaire. Les règles de transformation ne s'attacheraient initialement qu'à ce cadre. Sur cette base, il sera plus facilement possible d'identifier des erreurs manifestes (absence de la construction de base pour certaines entrées, attribution de règles de passivation à des constructions intransitives) et de contrôler l'extension du domaine d'application des transformations à certaines variantes constructionnelles du cadre canonique. L'appariement des arguments dans des constructions infinitives et diathétiques est un sujet difficile dont le traitement ne pourra se faire de façon satisfaisante qu'en articulant la syntaxe et la sémantique. LMF, en subordonnant la sous-catégorisation à la description du sens, ouvre, là aussi, une voie de modélisation. Enfin, l'analyse interne des constituants phrastiques enchâssés dans des constructions nominales (*de ce que P, le fait que P*) doit pouvoir se faire à différents niveaux de granularité, en fonction des besoins effectifs d'identification des constituants enchâssés.

Enfin, au vu d'un certain nombre de « bugs » qui peuplent le lexique actuellement diffusé, il ne semble pas inutile de rappeler que le choix d'un format de codage approprié, *i.e.* d'un

langage semi-structuré comme XML, aide considérablement lors de la phase de validation des données. Une simple validation par rapport à un schéma – fût-il propriétaire – permet de repérer certaines des anomalies relevées *supra*, mais aussi des valeurs erronées pour les identifiants, des pointeurs dont la cible n'est pas identifiable, la duplication d'une même valeur d'un attribut, ou l'incompatibilité entre des traits utilisés simultanément. Par ailleurs, la définition même d'un tel schéma implique de fait une réflexion sur la nature linguistique des traits et valeurs utilisées. Utiliser pour cela un cadre standardisé est évidemment un garant supplémentaire de cohérence, d'interopérabilité et d'une documentation pérenne.

Malheureusement, les lexiques actuellement disponibles pour le français ne relèvent pas d'une telle démarche : outre l'absence de standards en la matière, certains adoptent une théorie linguistique particulière, d'autres sont guidés par le besoin de comptabilité avec un analyseur en aval. La seule façon de mutualiser *a posteriori* ces ressources est de passer par un modèle de données (et son instanciation concrète) partagé. Dans cette optique, LMF a l'avantage de proposer un cadre à la fois standardisé et incrémental : dans un premier temps, une comparaison des catégories de données indispensables pour une description minimale des arguments syntaxiques pourrait aboutir à un référentiel stable, consigné dans le DCR. À partir de cet ensemble, il sera plus facile de déterminer les conditions d'unicité des arguments et d'aboutir à une description minimale des cadres dans une perspective extensionnelle, mettant de côté des questions théoriques plus difficiles, comme la représentation intensionnelle ou l'interfaçage avec les arguments sémantiques. Sur cette base, la comparaison et la mise en commun d'un noyau important des ressources paraît tout à fait envisageable dans un délai raisonnable. Dans un deuxième temps, les particularités de chaque lexique, souvent reliées à des points de vues théoriques particuliers, pourront être prises en compte, comme nous l'avons montré pour certaines caractéristiques du lexique-grammaire. Enfin, dans un scénario de mutualisation progressive des ressources, les questions de maintenance – et en particulier l'argument d'une certaine lourdeur dans la manipulation d'un format XML explicite – peuvent enfin être envisagées sous l'angle d'un investissement dans des infrastructures de gestion de lexiques pérennes.

Références

ANTONI-LAY M-H., FRANCOPOULO G., ZAYSSER L. (1994). « A generic model for reusable lexicons: The GENELEX project ». In *Literary and Linguistic Computing* 9 (1) : 47-54.

CREISSELS D. (1995). *Éléments de syntaxe générale*. PUF, Paris.

GARDENT C., GUILLAUME B., PERRIER G., FALK I. (2005). « Maurice Gross' Grammar Lexicon and Natural Language Processing ». In *Proceedings of the 2nd Language and Technology Conference,* Poznan.

GEYKEN A., BOYD-GRABER J. (2003). « Automatic classification of multi-word expressions in print dictionaries ». In *Linguisticae Investigationes* 26 (2) : 187-202.

GROSS M. (1975). *Méthodes en syntaxe.* Hermann, Paris.

IDE N., ROMARY L. (2004). « A Registry of Standard Data Categories for Linguistic Annotation ». In *Fourth International Conference on Language Resources and Evaluation* : 135-138.

KOENIG J.-P., MAUNER G., BIENVENUE B. (2003). « Arguments for adjuncts ». In *Cognition* 89 : 67-103.

MONTE G., FRANCOPOULO G. (2005). « Lexical Markup Framework (LMF) ». In *Working Draft ISO-24613* 7.

ROMARY L., SALMON-ALT S., FRANCOPOULO G. (2004). « Standards going concrete : from LMF to Morphalou ». In *Workshop on Electronic Dictionaries – Coling 2004*. Genève.

SAGOT B., CLÉMENT L., DE LA CLERGERIE E., BOULLIER P. (2006). « The Le*fff* 2 syntactic lexicon for French: architecture, acquisition, use ». In *LREC 06*. Genève.

SALMON-ALT S., AKROUT A., ROMARY L. (2005). « Proposals for a normalized representation of Standard Arabic full form lexica ». In *Second International Conference on Machine Intelligence (ACIDCA-ICMI 2005)*. Tozeur.

TESNIÈRE L. (1959). *Éléments de syntaxe structurale*. Klincksieck, Paris.

VAN DEN EYNDE K., MERTENS P. (2003). « La valence : l'approche pronominale et son application au lexique verbal ». In *French Language Studies* 13 (1) : 63-104.

Data Category Registry (DCR) – http://syntax.inist.fr

LEFFF2 – http://www.lefff.net

LEXIQUE-GRAMMAIRE – http://infolingu.univ-mlv.fr/

PROTON – http://bach.arts.kuleuven.be/PA/proton.html

SYNLEX – http://www.loria.fr/~gardent/ladl/content/resultats.php