CN RNIL N 7
2003-11-25

Secrétariat CN RNIL

Votre correspondant: **Tony HITTEMA**

Ligne directe: : 01 41 62 83 95
Télécopie : 01 41 62 90 33

E-mail : tony.hittema@afnor.fr

CN RNIL
Ressources Normalisées en Ingénierie de la Langue

Association

Française de

Normalisation

11 avenue Francis de Pressensé

93571 Saint-Denis La Plaine Cedex

France

Tél. : +33 (0)1 41 62 80 00

Fax : +33 (0)1 49 17 90 00

http://www.afnor.fr

TITRE : Proposition de norme des Lexiques pour le traitement automatique du langage

SOURCE : Gil FRANCOPOULO, INRIA/LORIA - Action Syntaxe

PROJET : -

STATUT : Avant-projet

ACTION : Pour information et discussion à la réunion du 2 décembre 2003

DIFFUSION : Membres de la commission

Association reconnue

d'utilité publique

Comité membre français

du CEN et de l'ISO

Siret 775 724 818 00015

Code NAF 751 ER

PROPOSITION DE NORME DES LEXIQUES POUR LE TRAITEMENT AUTOMATIQUE DU LANGAGE

Version-1.3 22 novembre 2003

Gil Francopoulo
INRIA/LORIA-ACTION SYNTAXE

Préambule

Le présent travail est réalisé dans le cadre de l'action SYNTAXE de l'INRIA, du projet RNTL - Outilex, de l'action Normalangue et du groupe de travail AFNOR : « lexiques pour le TAL ». Il est en relation avec l'initiative américaine ISO-LMF ainsi qu'avec les travaux sur les catégories de données du TC37/SC4 de l'ISO.

1- Introduction

Des ressources lexicales pour le traitement du langage (TAL) ont été créées, dans bien des cas en dehors du cadre des dictionnaires destinés à une édition imprimée. Mais il n'y a pas de format standard pour les structures décrites et le choix des catégories de données varie d'une ressource à l'autre. La connaissance inscrite dans de telles ressources est à la fois extensive et onéreuse à maintenir, et le besoin de fusionner ces ressources est à la fois fréquent et onéreux.
Des standards sont nécessaires afin de faciliter la création de telles ressources par des mécanismes de fusion ou par interopérabilité. Il s'agira ici de définir une norme de représentation des lexiques pour le traitement automatique du langage.

2- Périmètre de la spécification

L'objectif est de définir une norme de représentation des données lexicales dans les contextes de gestion et d'échange de dictionnaires.
Les données dont il est question ici sont exclusivement destinées aux applications du traitement automatique de la langue. La problématique des dictionnaires dits éditoriaux n'est pas abordée, elle est couverte par la norme ISO-1951 qui est d'ailleurs en cours de révision [Dérouin]. Les dictionnaires éditoriaux sont destinés à une lecture humaine, alors que nos données sont destinées à être exploitées par des programmes.
La tâche de définition d'un lexique pour le TAL est un petit peu plus difficile que le travail de définition d'un dictionnaire éditorial. Le lexicographe d'un dictionnaire éditorial suppose que son lecteur aura assez de connaissances pour interpréter et étendre l'information trouvée dans le dictionnaire. En revanche, cette connaissance implicite fait défaut à l'utilisateur du lexique pour le TAL puisque celui-ci est un programme. Ainsi, chaque mot doit être précisément et complètement décrit pour être d'un usage effectif par un programme de TAL.

Les applications du TAL que nous ciblons ne sont pas limitées. Il pourra s'agir de l'analyse, de la génération, de la traduction, de la correction, de l'extraction comme de l'indexation. Mais l'incidence du lexique ne se limite pas aux applications directes : dans la mesure où un certain nombre de terminologies d'entreprise sont couplées avec des applications de TAL. Il est en effet souhaitable que des passerelles soient définies entre la terminologie et le lexique. Pour éviter de parasiter la description du modèle, la passerelle Lexique-TMF est présentée en fin de document.

Les lexiques pourront être monolingues ou bien multilingues.
Les langues ciblées ne sont pas limitées.

3- Normes référencées

ISO 639-1:2002, Codes for the representation of names of languages – Part 1 : Alpha-2 Code.
ISO 8879:1986, (SGML) as extended by TC2 (ISO/IEC JTC 1/SC 34 N 029:1998-12-06) to allow for XML.
ISO 15924, Codes for the representation of names of scripts.
ISO 16642:2003, Computer applications in terminology – TMF (Terminological Markup Framework).
ISO 10646-1:2000, Unicode.
ISO CD12620-1, Model for description and procedures for maintenance of data category registries for language resources.
ISO NP12620-3, Data categories : electronic lexical terminological.

4- Crédits

La présente étude s'est inspirée des travaux suivants :

- Compilation des catégories de données fournies par les partenaires Outilex. Ce travail a été effectué par Sébastien Guérin du LORIA.
- Echanges de courriers électroniques de la liste de discussion « Lexiques pour le TAL » du site www.normalangue.org.
- La proposition de NWIP pour le TC37/SC4: Lexical Resource Markup Framework (LMF) par Monte George.
- De manière un peu plus générale, des travaux de l'équipe Papillon (notamment Christian Boitet et Mathieu Mangeot), d'Eric Laporte, de Gérard Huet, de Jean Senellart, de Genelex et de mon expérience d'une douzaine d'années chez Erli-LexiQuest.

5- Définitions

Nous distinguerons « lexicologie » qui est l'étude des mots, du terme « lexicographie » ou « dictionnairique » qui est l'activité de confection d'un dictionnaire. De ce fait, quand nous voudrons désigner un utilisateur du modèle et dans la mesure où notre intérêt porte sur la confection de lexiques, nous parlerons de lexicographe plutôt que de lexicologue. Nous n'utiliserons pas le terme de « linguiste » qui n'est pas assez précis.

Nous ne ferons pas de distinction entre « dictionnaire » et « lexique ». Notons que certains auteurs [Gaudin] distinguent le dictionnaire comme étant un recueil où chaque mot possède une définition contrairement à un lexique dans lequel la définition n'est pas obligatoire.

Nous appelons un trait morphologique, un axe de valeurs associées à la morphologie d'un mot. Nous appelons une valeur de trait, une constante particulière de ce trait. Ainsi, le trait morphologique de nombre de « maisons » aura pour valeur /plural/.
Les traits morphologiques et leurs valeurs sont définis par le « Data Category Registry » de la norme ISO-12620.
Le plus souvent nous ne parlerons pas du trait morphologique de manière isolée mais de la combinaison de différents traits morphologiques. Une telle combinaison sera par exemple : /number/ + /gender/. Nous parlerons aussi de combinaisons de valeurs de traits. Ce sera par exemple : /plural/ + /masculine/.

Nous éviterons prudemment d'employer le terme « lexème » qui possède des sens sensiblement différents selon les auteurs. Par exemple dans la théorie Sens-Texte, un lexème est une sorte de « lexie » qui s'oppose à un « phrasème », ou locution. C'est le sens d'un mot simple. Au contraire, dans LMF il n'y a pas cette distinction, le lexème inclut le « phrasème ».

Toutes les chaînes de caractères sont exprimées en ISO10646-1:2000 (i.e. Unicode) donc nous ne préciserons pas le codage dans la description du modèle.

Les références au registre des catégories de données ISO 12620 seront exprimées entre deux slash. Ce sera par exemple : /gender/.

Dans les diagrammes, nous distinguerons les relations et les listes. Un élément pourra être en relation avec un élément B, cela signifiera simplement que les éléments sont liés. Dans le diagramme, la représentation sera une flèche dotée d'une cardinalité. Dans la liste, nous avons la notion d'ordre, ainsi les éléments figurant dans la liste pourront être référencés par leur numéro d'ordre. La liste sera représentée par une boîte.

6- Etat de l'art

Depuis une vingtaine d'années un grand nombre de modèles de dictionnaires ont été définis et utilisés, avec plus ou moins de généralité. Un nombre plus restreint a été publié, et nous nous limiterons à ces derniers.
Nous pouvons schématiquement décrire ces modèles en fonction de leur filiation qui se résume à six familles généalogiques.

a) La famille de Princeton
Le modèle original est celui du lexique WordNet en version anglo-américaine créé par l'Université de Princeton. Les modèles dérivés sont EuroWordNet pour les langues de l'Europe de l'Ouest, ItalWordNet pour l'italien, IndoWordNet pour l'Asie et BalkaNet pour les langues de l'Europe de l'Est.
La version anglo-américaine de WordNet est certainement le dictionnaire le plus répandu car il est gratuit et porte sur la langue la plus répandue du domaine.

b) Les modèles Européens complexes
Le modèle original est Genelex. Les modèles dérivés sont Eagles, Parole, Simple, Isle et Mile. Le modèle Relex de l'IGM fait partie de cette famille.
Ces travaux ont fortement influencé la pratique du TAL en Europe. Ces modèles sont puissants, mais le principal reproche à leur faire est que ce sont des modèles complexes qui ne sont pas simplifiables. Ils sont le fait de consortiums qui, pour satisfaire tout les partenaires ont produit l'union des mécanismes de représentation. De ce fait, un grand nombre d'acteurs en Europe n'implémente qu'une petite partie de ces modèles sans prendre en compte la totalité.

c) Les modèles Européens simples
Ce sont BDLex, Celex, Multex et « Multex goes East ». Notons que ces modèles n'ont pas forcément de liens de filiation entre eux : ils partagent le fait qu'ils sont simples et traitent principalement de morphologie.

d) Le modèle EDR
Il s'agit du modèle du consortium japonais EDR dont les travaux sont repris actuellement par le CRL.
C'est un modèle bilingue spécifiquement destiné au couple japonais-anglais.

e) La famille des modèles de Mel'cuk
Le modèle original est le DEC (Dictionnaire explicatif et combinatoire) qui est un peu particulier dans la mesure où le dictionnaire associé n'a pas été utilisé pour le TAL. En revanche, les modèles dérivés comme DiCo ou Papillon sont destinés au TAL.
La théorie sous-jacente, qui est la théorie Sens-Texte, fait de cette famille de dictionnaires un exemple significatif de travail lexicographique. La pratique en question se distingue par un jeu de critères méthodologiques particulièrement rigoureux.

f) La famille TEI
Ces modèles se fondent sur les directives TEI. Ce sont ALLEX (African languages lexicons) et CJKE (Chineese, japaneese, Korean and English). Notons que leur utilisation dans une perspective TAL n'est pas avérée.

7- Problématique

Du fait de la diversité des modèles, l'échange de données n'est pas très facile, surtout entre des lexiques de familles différentes. La fusion de dictionnaires reste une opération très complexe.
Il n'est pas très facile non plus de faire cohabiter des programmes qui opèrent sur des modèles de lexiques différents.

Et pour certains modèles, la structure n'est pas très bien définie. Certains modèles ont quelques difficultés à représenter les informations linguistiques que **la langue nous force à décrire** : le phénomène est particulièrement criant pour la description des mots composés et les opérations de transfert verbal pour la traduction.

8- Critères imposés

L'orientation que nous adoptons est régie par les critères que nous nous imposons.

a) Critère de simplicité
Le modèle doit être simple pour un lexicographe qui désire la simplicité.
Le modèle doit être puissant pour un lexicographe qui veut de l'expressivité tout en admettant un peu de complexité.
b) Critère de représentativité
Le modèle doit être capable de représenter, dans la mesure du possible, les dictionnaires existants. Si tel n'est pas le cas, l'information problématique doit pouvoir être détectée et isolée.
c) Critère de distinction par rapport aux autres modèles
Le modèle ne doit pas constituer une septième famille mais doit à la fois s'inspirer de tous les modèles existants et en être une représentation pivot.

9- Modèle proposé

a) Choix

a-1) nombre de couches
Concernant la structure monolingue du modèle, nous avons deux possibilités. La première consiste à définir un modèle avec deux couches : la morphologie et la sémantique. La seconde consiste à intercaler une information syntaxique entre la morphologie du mot et ses différents sens. Cela produit un modèle à trois couches : la morphologie, la syntaxe et la sémantique.
Par expérience, nous savons que les modèles à trois couches ne peuvent respecter le critère de simplicité car l'information syntaxique est un passage obligé entre la morphologie et la sémantique, et complexifie chaque entrées lexicale, même si l'on ne désire pas décrire spécialement le comportement syntaxique du mot. Pour cette raison, nous choisissons une modèle à deux couches. Si nous avons besoin de représenter les informations syntaxiques nous les projèterons sur les sens du mots, mais cette projection sera optionnelle.

a-2) absence vs présence de la morphologie
Certains acteurs du domaine ne désirent pas faire figurer le calcul de la morphologie associée à chaque mot au prétexte que c'est trop soi-disant trop complexe ou bien qu'il n'y a pas consensus. L'argument qui milite contre cette position est que tous les dictionnaires pour le TAL possèdent l'information morphologique : il est en effet impossible d'utiliser un dictionnaire si celui-ci n'est pas capable d'associer une forme lemmatisée à une ou plusieurs formes fléchies, ceci est vrai pour les mots simples comme pour les mots composés.
Il faut considérer que l'information morphologique fait partie du dictionnaire au même titre que les entrées, si le modèle en tient compte, les possibilités d'échanges de données seront beaucoup plus importantes que si tel n'est pas le cas.

Imaginons le contexte suivant : supposons que nous ayons des logiciels et des lexiques qui fonctionnent en N langues et qu'il s'agisse d'ajouter une langue par acquisition via un format normalisé. Si la morphologie de la langue est relativement simple comme peut être l'anglais, le problème n'est très important. Mais si la morphologie de la langue en question est complexe comme le français (51 formes pour chaque verbe) ou le hongrois (238 formes pour chaque nom), il sera très appréciable de disposer de la morphologie avec les entrées plutôt que d'avoir à créer les paradigmes de flexion par ailleurs.

Il est vrai qu'il n'y a pas consensus à ce propos et l'on peut schématiser la situation en trois orientations :

- le code fait référence à un automate comme le pratique l'IGM,
- le code est une description symbolique comme Genelex,
- le code est une référence à un programme compilé opaque.

Pour ceux qui pratiquent la dernière orientation, il n'y aucun espoir de modélisation ou d'échange de données.
Nous prenons le parti suivant :

- la description est optionnelle,
- la description est symbolique et peut être compilée sous forme d'automate.

Pour la morphologie des mots simples, la description repose sur une liste de radicaux qui est modifiable par une séquence de cinq opérateurs linguistiques (c.f. détails dans le système de la morphologie).
Pour la morphologie des mots composés, la description repose sur les combinateurs de traits morphologiques en s'inspirant du modèle Eagles mais en l'améliorant.

a-3) représentation graphique vs représentation phonétique
La forme lemmatisée graphique est obligatoire, en revanche sa représentation phonétique est optionnelle. Si le lexicographe décide de décrire la morphologie graphique d'un mot, il n'est pas obligé de représenter la phonétique du mot. De même, si le lexicographe décide de représenter la morphologie phonétique du mot, il n'est pas obligé de représenter son équivalent graphique.

a-4) multilinguisme
Le multilinguisme est privilégié par rapport au bilinguisme. Le mécanisme des axies (e.g. Papillon) est préféré à la définition d'un lien bilingue (e.g. EDR).

b) Structure
L'esprit général est le suivant :

Pour chaque mot, nous avons un squelette et des systèmes périphériques.
Le squelette est rigide et obligatoire : il est simple.
Les systèmes périphériques sont au contraire souples et optionnels : ils sont puissants.

Le squelette est composé :

- du système de l'entrée lexicale,
- du système du sens.

Nous appèlerons un mot bien formé, un mot dont le squelette est complètement décrit.

Les systèmes périphériques sont :

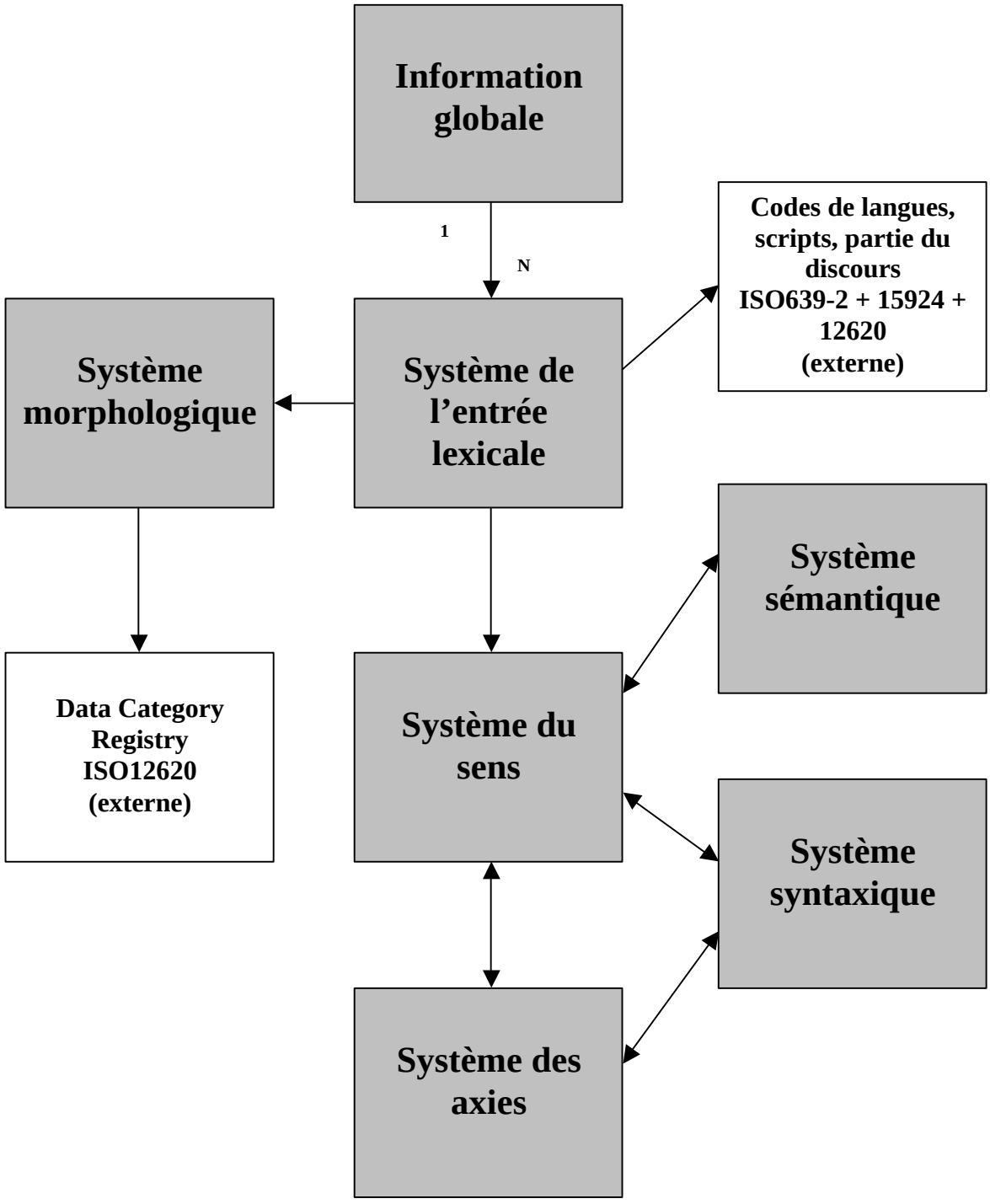
- la morphologie,
- la syntaxe,
- la sémantique,
- les axies.

Afin de présenter la structure progressivement, nous commençons par présenter l'architecture générale puis, en détail, nous décrirons chaque sous-système.

c) Architecture générale

L'information globale est un élément **unique** qui permet de d'enregistrer, entre autres choses, le nom, la version, les mentions de copyright.

Par convention, dans le diagramme, seuls les éléments grisés appartiennent au modèle. Les informations dans des boîtes blanches sont prises dans d'autres normes.



d) Détail de chaque sous-système

d-1) Le système de l'entrée lexicale

L'entrée lexicale sert à représenter le mot en tant qu'entité morphologique, que l'on décrive ou non sa morphologie complète. C'est le 'Morphological Unit' du modèle Eagles. C'est le signifiant de Saussure.
Il y a autant d'entrées lexicales qu'il y a de mots dans le dictionnaire.

L'entrée contient les attributs obligatoires :

- /identifier/,
- le code de la langue qui est une valeur prise dans la norme ISO639-2,
- le code du système de script qui est une valeur prise dans la norme ISO15924.
- la catégorie et sous-catégorie grammaticale (partie du discours), qui sont deux valeurs prises dans le registre de catégorie de données de la norme ISO12620.
- la graphie lemmatisée qui est une chaîne de caractères.
- si le mot est (ou non) autonome.

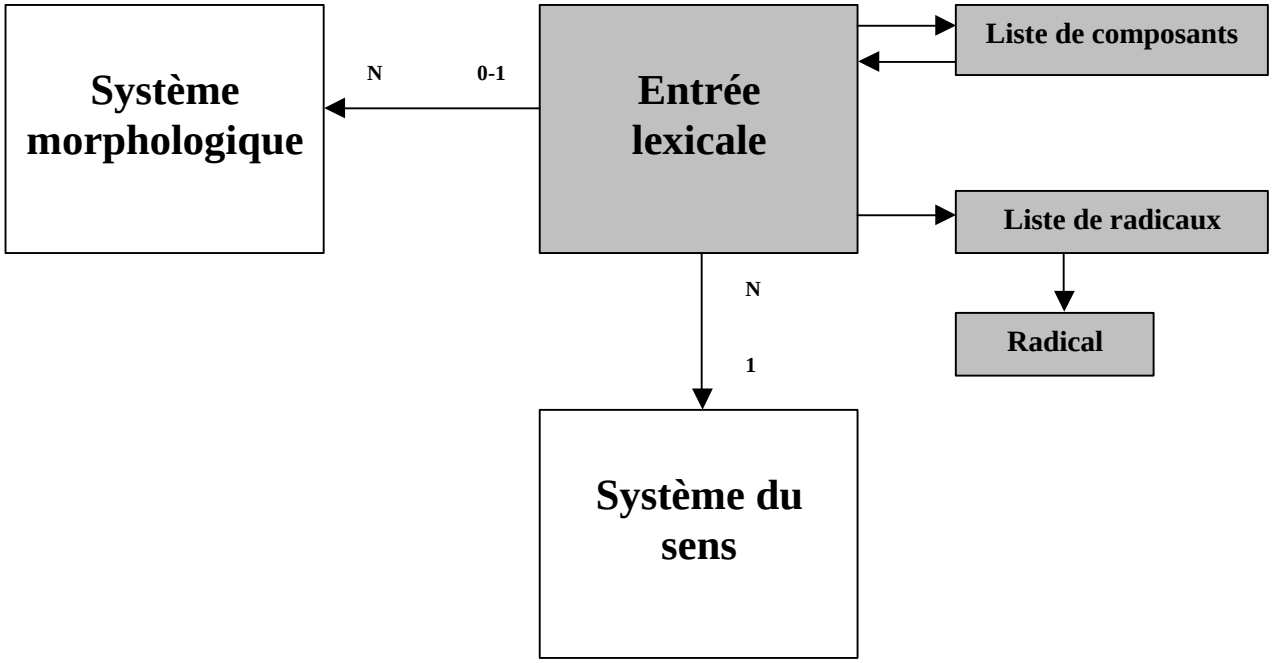
Les attributs optionnels sont les suivants :

- la représentation phonétique lemmatisée qui est une chaîne de caractères.
- la composition. C'est la liste des composants qui sont d'autres entrées lexicales. Quand la liste est vide, l'entrée lexicale est celle d'un mot simple.
- les radicaux. C'est la liste sur laquelle repose la morphologie flexionnelle (voir le système morphologique). Chaque radical porte une chaîne de caractères optionnelle pour la graphie et une chaîne optionnelle pour la phonétique. Au moins une des deux chaînes doit être renseignée. Un radical n'est pas partagé par plusieurs entrées lexicales.

Dans ses relations avec les systèmes voisins :

- l'entrée lexicale possède zéro ou un lien vers une description morphologie. Concernant l'absence de lien, ce n'est pas que la description n'existe pas en langue, c'est que l'on n'impose pas au lexicographe.
- l'entrée possède zéro à N sens.

Par convention, seuls les éléments grisés appartiennent au système de l'entrée lexicale.



d-2) Le système du sens

C'est le système du modèle LMF. Dans la théorie Sens-Texte, l'élément s'appelle « lexie ». Pour Eagles, l'élément s'appelle « Semantic Unit ». Pour Saussure, c'est le signifié. En français, c'est une acception.

Un sens n'est relié qu'à une seule entrée lexicale.

Le sens contient l'attribut obligatoire :

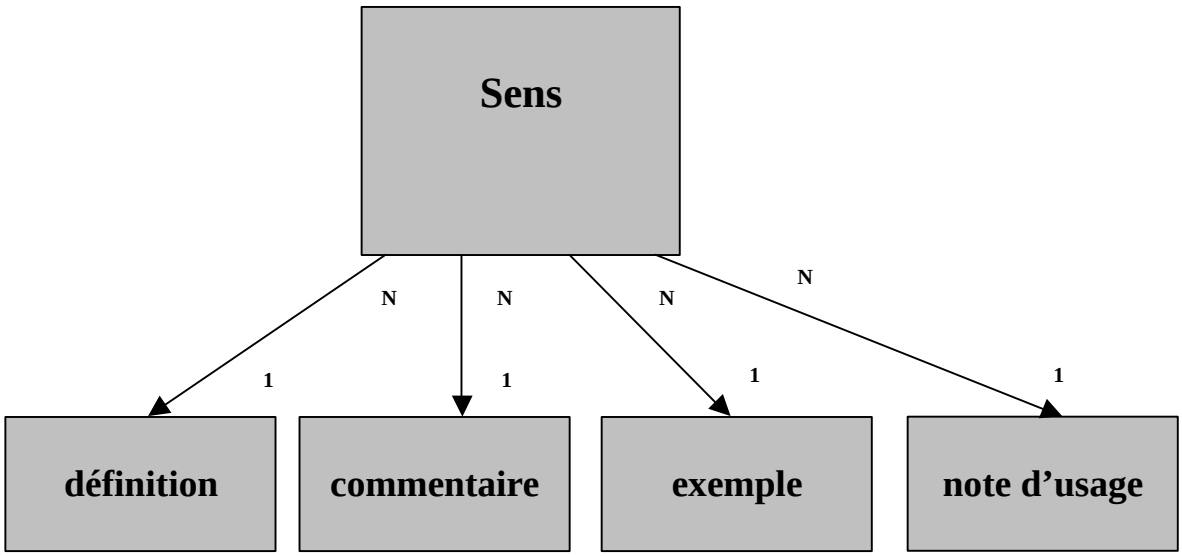
- /identifier/,
- la KeyForm pour la représentation externe, par exemple : 'souris (animal)', exprimée sous forme d'une chaîne de caractères.

Les éléments additionnels sont :

- la définition sous forme d'un texte,
- le commentaire,
- l'exemple,
- la note d'usage.

Ces éléments peuvent être multiples et sont optionnels.
Chacun d'eux porte deux attributs :

- le contenu, sous forme d'une chaîne de caractères.
- la langue d'expression qui n'est pas nécessairement la langue du mot. La valeur est prise dans la norme ISO 639-2.



d-3) Le système morphologique

d-3.1 Introduction

La fonction du système est de produire toutes les paires : combinaison de valeurs de traits morphologiques / formes fléchies. Par exemple, pour l'entrée 'go', une de ces paires sera : (/third person/ + /singular/ + /present/) / 'goes'.

Le système traite les mots simples et multiples.
Le système fonctionne en génération aussi bien qu'en analyse.
Le système traite l'information graphique et phonétique.

d-3.2 Le mot simple
Le système est celui d'Eagles qui a été modifié pour être rendu plus puissant.

Le mécanisme pour décrire la morphologie d'un mot simple est le suivant :

- on fait référence à la forme lemmatisée ou à une liste de radicaux qui est attachée à l'entrée lexicale. La référence zéro signifie la forme lemmatisée. Un entier supérieur à un est le rang du radical.
- on spécifie des opérateurs que l'on applique à la chaîne obtenue au point précédent. Il y a entre zéro et cinq opérateurs qui sont appliqués dans l'ordre dans lequel ils sont présentés dans la séquence suivante :
  - opérateur-1 : retirer N caractères en début,
  - opérateur-2 : ajouter une chaîne en début,
  - opérateur-3 : déplacer N caractères depuis la position X jusqu'à la position Y,
  - opérateur-4 : enlever N caractères en fin,
  - opérateur-5 : ajouter une chaîne en fin.

  Chaque opérateur ne s'applique qu'une seule fois.

Le paradigme de flexion porte les attributs suivants :

- /identifier/.
- exemple.
- partie du discours auquel il est destiné. C'est une référence à un registre de catégories de données.
- si le paradigme s'applique à la représentation graphique,
- si le paradigme s'applique à la représentation phonétique.

Le combinateur de traits morphologique ne porte aucun attribut propre, seulement des liens. C'est un objet non-linguistique qui ne sert qu'à relier des éléments.

Le composeur porte les traits suivants :

- s'il est simple ou composé,
- le rang.

Le calculateur de formes fléchies porte un double jeu d'opérateurs : un jeu pour le calcul graphique et un jeu pour le calcul phonétique.
Le système est certainement plus facile à comprendre en consultant le chapitre des exemples de mots.

d-3.3 Le mot multiple
Le système retenu est une combinaison des mécanismes de composition morphologique et syntaxique de Genelex. En fait, on est plus ou moins obligé de faire intervenir des éléments de syntaxe même si celle-ci reste locale.

On peut définir le mot multiple selon deux vues :
1) L'aspect intrinsèque :
Si le mécanisme des mots simples combinent des chaînes pour former un mot, celui des mots multiples combine des mots pour en former un autre.
2) L'aspect extrinsèque
Le mot multiple est une séquence de mots qui se comporte comme un unité simple à un certain niveau de l'analyse linguistique [Calzolari].

Un mot multiple est soit :

- un mot composé continu comme : « pomme de terre ». Le mot n'est pas nécessairement un nom, il peut être de n'importe quelle catégorie grammaticale.
- un mot aggluti né, pour les langues agglutinantes.

- un mot discontinu comme « passer en revue ». On appelle un mot discontinu, un mot qui peut admettre une insertion sans que celle-ci soit obligatoire. On pourra former : « passer N1 en revue » tout aussi bien que « passer en revue N1 » sans qu'il soit nécessaire d'enregistrer deux entrées lexicales.

Un mot multiple peut être une construction à verbe support comme « faire une acquisition ».

Un mot multiple est constitué de composants qui sont ou non autonomes. Ainsi, dans « au fur et à mesure », le composant « fur » n'a pas d'existence en tant que mot isolé, on l'appelle alors un mot non-autonome.

Le mécanisme s'applique pour former les agglutinés des langues agglutinantes. En fait, on considère qu'un mot agglutiné est un mot composé qui ne comporte aucun séparateur graphique.

Le mécanisme s'applique récursivement : un mot multiple peut être constitué de composants qui sont eux-même des mots multiples. Par exemple : l'adjectif composé « à haute tension » peut former le nom composé « ligne à haute tension ». De même, un mot multiple peut comporter des composants qui sont des agglutinés.

Le système combine trois types de descriptions :

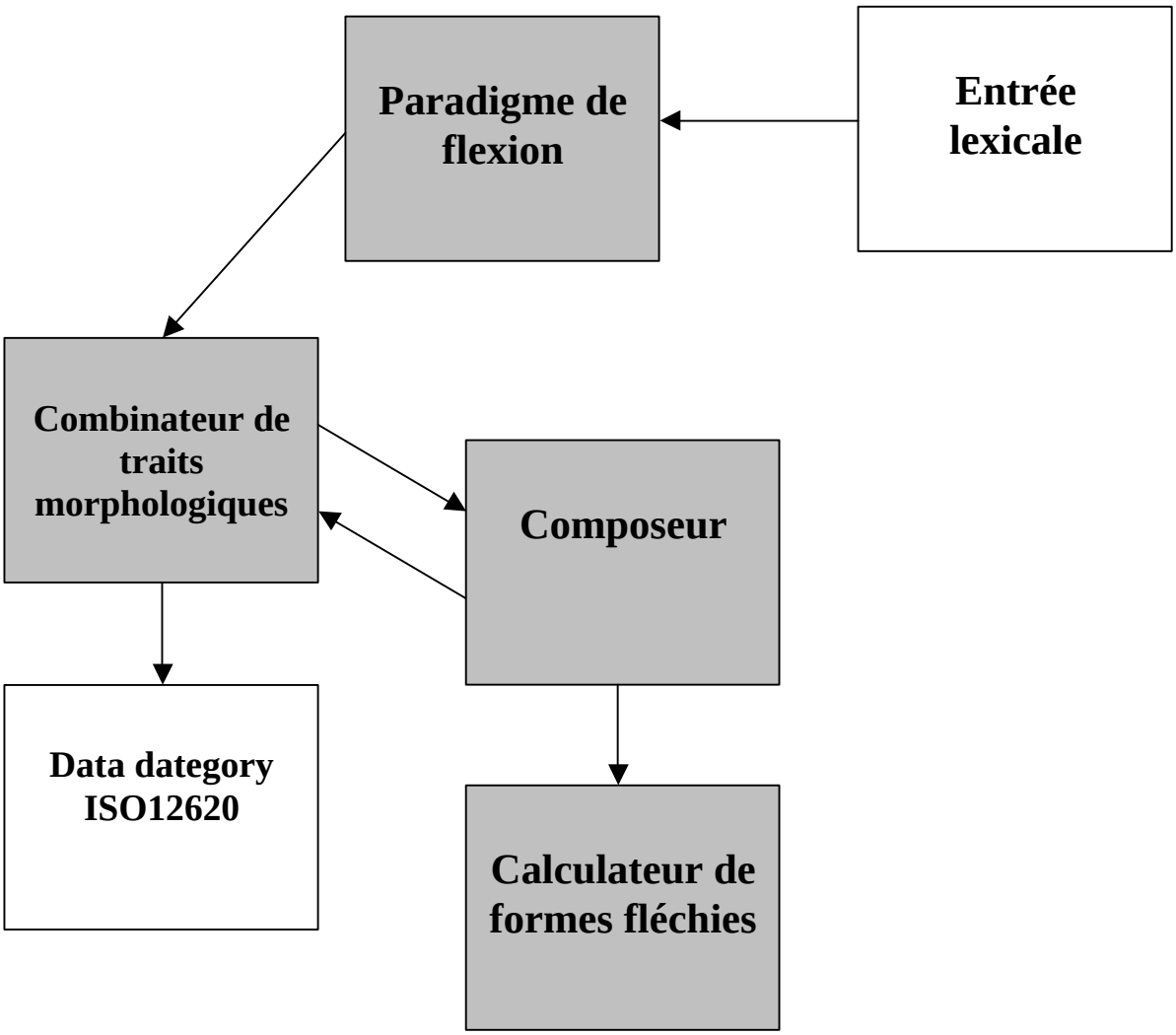
- une information qui porte sur plus d'un mot. C'est un codage qui peut porter sur la totalité du mot : par exemple, si le composé admet ou non un ordre libre de ses composants.
- les insertions possibles en précisant leur type et position. Ce sera utile pour décrire « passer N1 en revue ».
- une information qui porte sur chacun des composants en précisant :
  - la combinaison de traits morphologiques. C'est ce qui va déterminer le calcul des formes fléchies à partie des composants.
  - la référence au lemme ou radical. C'est un entier.
  - le séparateur graphique, par exemple, un tiret, un espace ou la possibilité d'avoir les deux. La valeur « jointure » permet de représenter l'agglutination. Par convention, le séparateur s'applique après le composant.
  - un code libre, afin par exemple de spécifier un changement de casse. Dans certaines langues agglutinantes, le composant autonome commence par une majuscule qui devient minuscule en agglutination.

Notons que nous ne représentons pas deux types d'information :

- la désignation de la tête.
- l'accord entre une sous-partie des composants. En fait, on l'exprime plus ou moins : on n'exprime pas explicitement qu'un composant s'accorde avec un autre composant, mais l'on spécifie que le composant s'accorde de telle ou telle manière avec le composé.

En français, pour les mots composés continus, nous avons les structures régulières qui sont au nombre de 9 : Adj-N, N-Adj, NàN, NàGN, NdeN, NdeGN, N-N , V-N, Prép-N [Silberztein], [Gaudin]. En définissant ces neuf paradigmes de flexion composée, nous couvrons l'écrasante majorité des mots composés du français. La tâche n'est pas finie pour autant car il reste un certain nombre de mots qui sont particuliers et qui nécessitent une description spécifique.

Le système est certainement plus facile à comprendre en consultant le chapitre des exemples de mots.



d-4) Le système sémantique

Il s'agit de représenter les relations sémantiques et les traits sémantiques.

Une relation permet de qualifier le lien qu'entretient deux sens. Ces derniers doivent appartenir à la même langue, en effet, les mécanismes de traduction sont couverts par un autre système : le système des axies.

Un trait est simplement une information attachée à un sens. Une trait est un cas particulier de la relation : c'est une relation qui n'a pas de cible.

La relation permet de représenter la dérivation sémantique, la synonymie, la parasynonymie des dictionnaire éditoriaux. Il peut tout aussi bien représenter les fonctions lexicales du DEC. Les liens de synonymie permettent aussi de décrire les synsets (synonyms in a set) de WordNet.

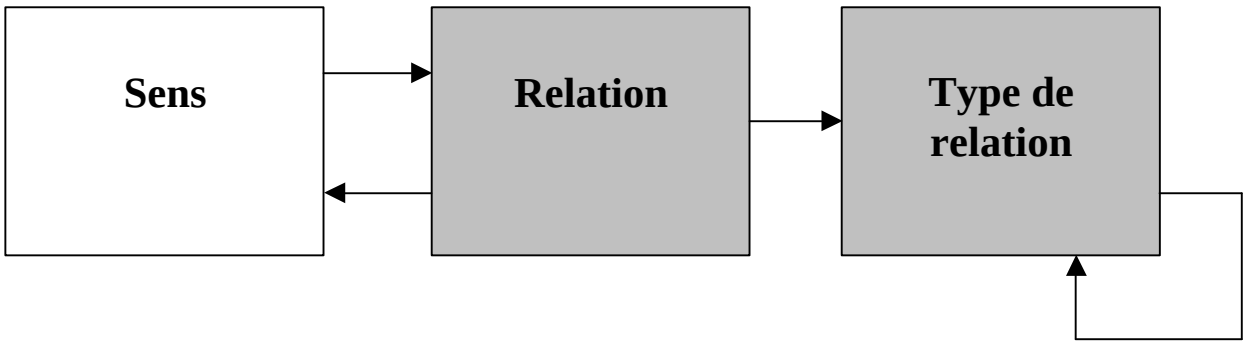
Le trait permet de représenter par exemple, le mécanisme du champ sémantique du DEC.

Plus précisément, le système comprend deux éléments : la relation qui est l'instance qui relie les sens et le type de relation qui permet de qualifier de manière factorisée la relation. Les types de relation peuvent entretenir entre eux des liens d'héritage.

Le type de relation dispose des attributs suivants :

- /identifier/,
- nom qui est une chaîne,
- commentaire qui est une chaîne,
- exemple qui est une chaîne,
- lien d'héritage.

Par convention, les éléments du système sémantique sont grisés.

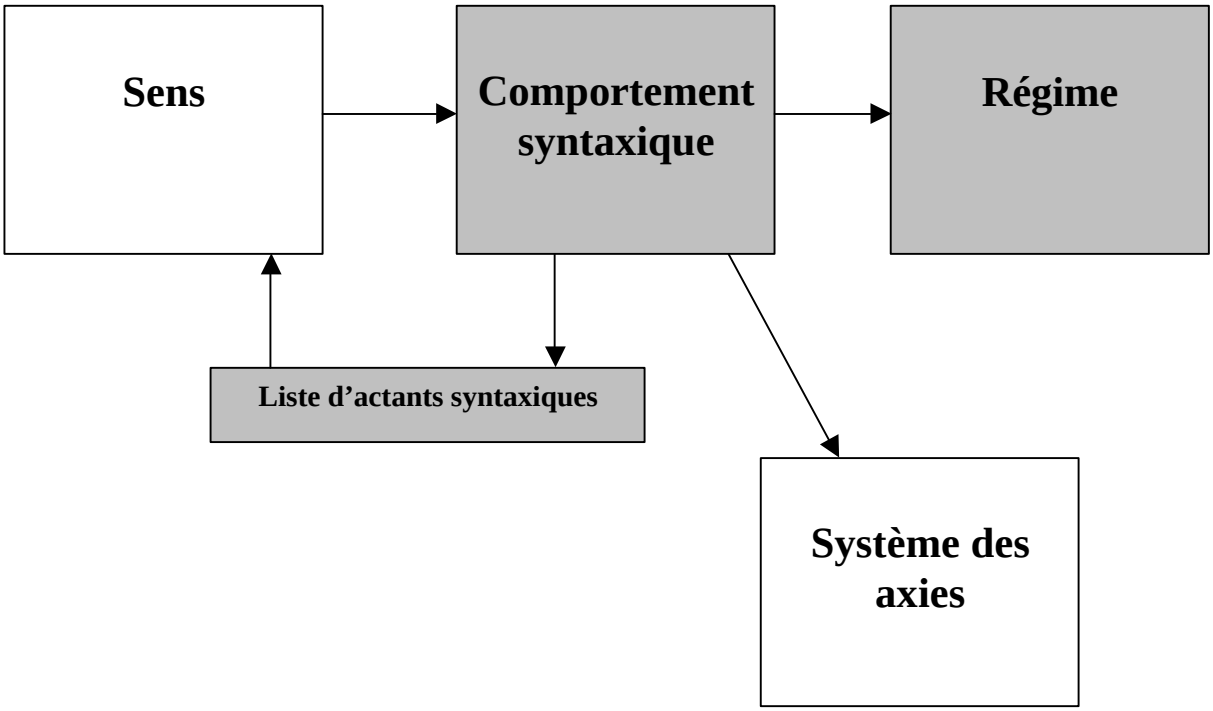


d-5) Le système syntaxique

Il s'inspire des régimes du DEC. C'est un système plus simple que les FramSet d'Eagles dans la mesure où les actants syntaxiques ne sont pas des unités syntaxiques mais des sens : il n'y a d'ailleurs pas d'unités syntaxiques dans notre modèle.

Un sens peut avoir plusieurs comportements syntaxiques. L'élément « comportement syntaxique » est simplement un connecteur entre des éléments. L'information linguistique importante figure sur le régime.

Le système permet de rendre compte du comportement syntaxique du sens. Le système est surtout destiné aux verbes, aux noms prédicatifs et aux adjectifs.



d-6) Le système des axies
Le mécanisme s'inspire du modèle Papillon.

Dans un dictionnaire bilingue, nous avons besoin d'un lien pour traduire un sens en un autre et on pourrait imaginer qu'il suffit d'un simple lien entre deux sens.

En fait, il y a deux types de problèmes :
a) **Premier problème**
Dans certains cas, cela ne fonctionne pas très bien parce que la finesse de la langue source n'est pas la même que celle de la langue cible. Ainsi, pour traduire le français « fleuve » (rivière qui se jette dans la mer) en anglais, nous arrivons pas à être aussi précis parce que le mot (donc le sens) n'existe pas dans la langue cible. Une solution consiste à créer un objet intermédiaire (que certains appellent une « cheville ») et d'indiquer une spécialisation par rapport au lien de traduction « rivière » vers l'anglais « river ».
b) **Deuxième problème**
Si cette stratégie est viable pour deux langues, elle est intenable pour un nombre de langues plus important comme quatre ou cinq tout bonnement parce que le nombre de liens explose.

Pour éviter ces problèmes, nous représentons les traductions via un objet intermédiaire que l'on appelle « axie ». C'est une structure pivot qui met en relation des éléments appartenant à des dictionnaires de langues différentes.
C'est une structure qui n'a pas lieu d'exister dans un dictionnaire monolingue.

D'autre part, en ce qui concerne les TAL multilingues, il y a deux écoles : celle du transfert et celle du pivot. Le transfert opère en syntaxe et le pivot en sémantique. Le transfert consiste à traduire en se fondant sur l'information syntaxique d'une langue pour aboutir à une structure syntaxique de la langue cible. L'approche via un pivot consiste à retrouver un élément qui ne dépend pas de la langue (aka pivot interlingua) et ensuite à engendrer la phrase dans la langue cible via un programme de génération. Notons que la majorité des outils de traduction (ou d'aide à la traduction) commerciaux se fondent sur l'approche transfert. Mais, on ne peut négliger les partisans de l'approche pivot qui se focalisent sur des niches techniques ou qui sont associés à des mécanismes de représentation interlingua. De plus, la traduction n'est pas la seule application d'un lexique multilingue : la recherche trans-linguistique d'information (i.e. Cross-Lingual Information Retrieval) est un domaine important, même si il est moins visible [Peters]. Et dans ce domaine, la répartition transfert vs pivot est moins tranchée.

De ce fait, le modèle de lexique doit permettre de pratiquer les deux approches. Notons qu'il est techniquement envisageable de combiner les deux approches, même si la complexité accrue serait problématique.

On distingue :

1) L'axie basique qui relie deux sens de deux langues différentes.
L'axie basique sert à implémenter l'approche pivot. Elle peut servir aussi pour l'approche par transfert, dans les situations où l'on dispose de la traduction sans avoir quoi que ce soit à exprimer sur la syntaxe du sens.
L'axie permet de traduire des mots qui n'ont pas nécessairement le même statut d'une langue à l'autre. Ainsi, dans la langue source, nous pourrons avoir un mot simple et qui se traduira par un mot composé dans la langue cible.
Les axies entre elles peuvent être décrites les unes par rapport aux autres via une relation d'axie.
La relation d'axie porte trois attributs optionnels :

- un intitulé qui est une chaîne de caractères.
- le nom d'un système descriptif externe.
- la référence dans le système descriptif externe.

L'intitulé permet de coder des relations interlingua simples comme le raffinement de « fleuve » comparativement à « rivière » et « river ». Mais il n'a pas pour objectif de coder un système complexe de représentation de connaissances : pour cela, il est préférable d'utiliser un système cohérent, complet et surtout conçu pour cela. Un bon candidat est UNL, éventuellement associé à des fonctions lexicales pour représenter des choses un peu délicates comme les collocations à verbe support [Boguslavsky] dans un système pivot.

2) **L'axie syntaxique** qui permet de réaliser le transfert.
On peut rendre compte des phénomènes d'inversion d'actants syntaxiques comme :

« FR :Elle me manque » => « EN : I miss her »

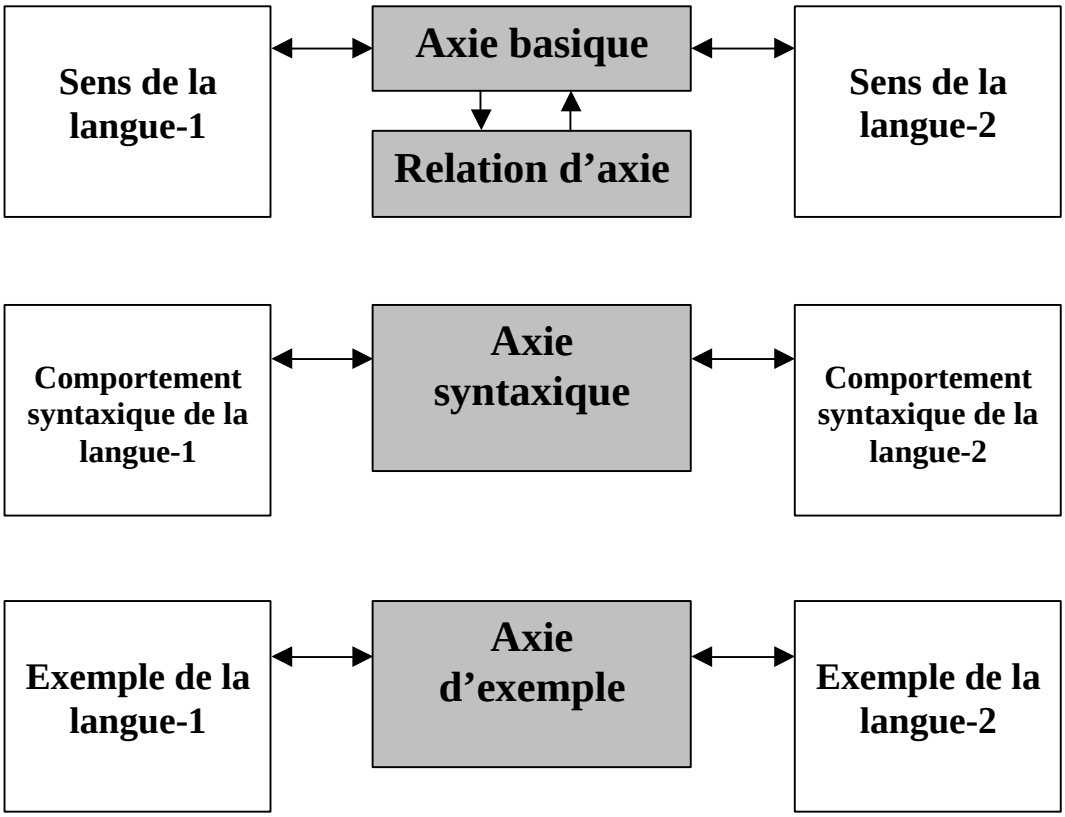
Dans la mesure où une entrée lexicale peut être une construction à verbe support, on peut rendre compte des traductions qui font passer d'un verbe plein (dans la langue-1) à un verbe support (dans la langue-2) comme :

"FR: Marie rêve" => "JP: Marie wa yume wo miru".

Cette axie ne porte aucun attribut.

3) **L'axie d'exemple** qui permet de documenter la traduction des exemples. Les exemples de sens peuvent être traduits mais ce n'est pas une obligation. Dans la mesure où on peut avoir plusieurs exemples pour un seul sens, nous avons besoin d'un mécanisme de d'association d'un exemple d'une langue en un exemple d'une autre langue. Quelques fois les exemples font référence à des objets culturels et il est nécessaire de transposer les référence d'une culture à l'autre. En adaptant l'exemple de [Boitet], en français, on dirait : « Pour mes voyages, je fais confiance à l'automobile club de France », alors qu'un américain dirait : « For travelling, I trust the American Automobile Association ».
Cette axie ne porte aucun attribut.

Par convention, les éléments du système des axies sont grisés.

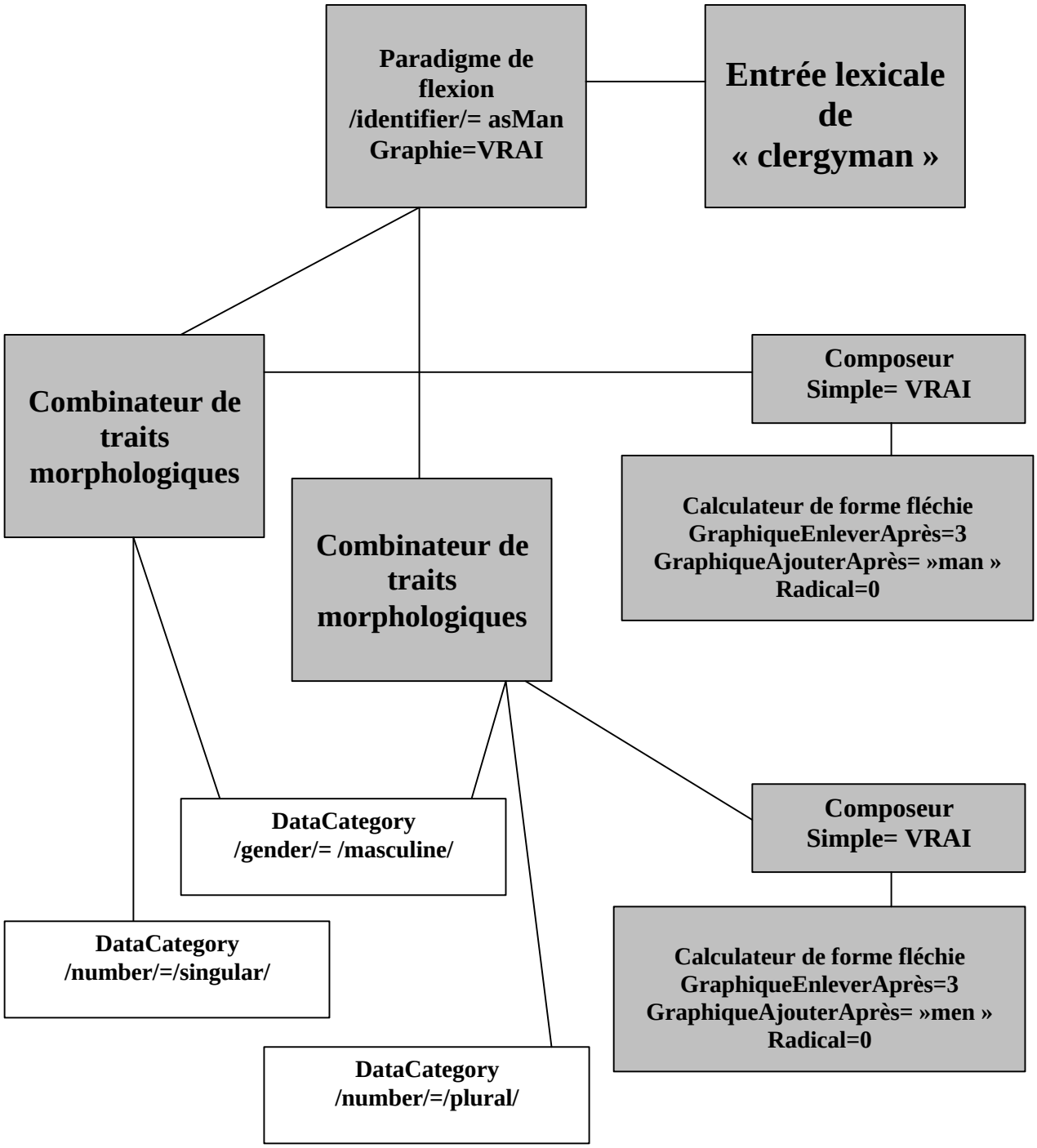


e) Exemples de mots

e-1) Le mot anglais « clergyman »

C'est un exemple d'application du paradigme à un mot simple. Le singulier est « clergyman » et le pluriel est « clergymen ». Le mode de flexion s'appelle « asMan ». Comme la morphologie de la langue anglaise est relativement simple, nous prenons le parti de rester simple. Donc, nous ne gérons aucun radical et procédons par référence à la forme lemmatisée. La valeur de l'attribut « radical » vaut donc zéro.

Prenons la convention que les éléments grisés font partie du modèle et que les éléments blancs sont externes.



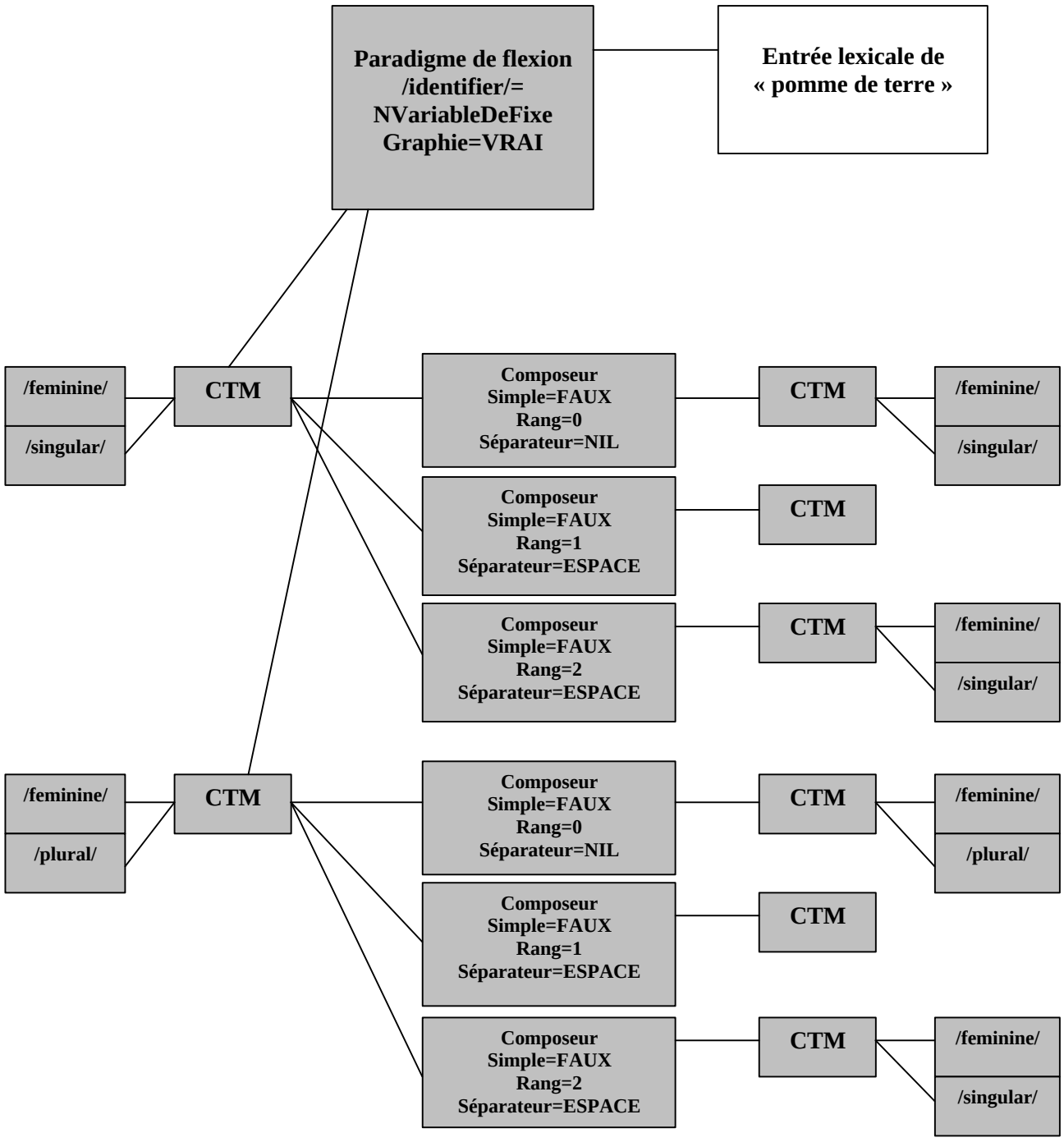
e-2) Le mot français « pomme de terre »

Les formes fléchies sont calculées depuis les composants en faisant référence aux combinaisons des traits morphologiques de chacun des composants.

Par convention, et pour gagner de la place sur le diagramme, « Combinateur de traits morphologiques » est abrégé en CTM.

Le singulier de « pomme de terre » est « pomme de terre ». Le pluriel est « pommes de terre ». C'est un comportement très commun pour une forme NdeN d'avoir une variation sur la seule tête du mot composé avec un figement sur la partie modifiante.

Les CTM sur la partie gauche représentent le nombre et le genre du composé. Les CTM de la partie droite représentent le nombre et le genre des composants. La préposition « de » n'est pas liée à une DataDategory car elle n'a aucun trait morphologique.



e) synthèse

La lexicographie éditoriale a pour tradition de considérer qu'un dictionnaire se définit selon trois critères [Rey-Debove] :

- la nomenclature : quels sont les mots du dictionnaire ?
- la microstructure : l'organisation d'un article qui se répète de façon systématique,
- la macrostructure : l'organisation de l'ensemble du dictionnaire.

Pour notre modèle, c'est un peu différent car nous avons un axe supplémentaire qui est celui des constantes linguistiques. Ce sont des éléments qui sont caractéristiques d'une langue et qui sont utilisés par un grand nombre d'éléments.

La microstructure est constituée de l'entrée lexicale et du sens.

La macrostructure est constituée des systèmes sémantiques (sauf type de relation), syntaxiques (sauf régime) et des axies.

Les constantes linguistiques sont les éléments de la morphologie, le type de relation, le régime ainsi que l'information globale.

10- Connexion avec TMF

TMF est un méta-modèle dédié à la terminologie et décrit par la norme ISO-16642 [Romary].

Il est souhaitable qu'un utilisateur de TMF puisse associer des outils de TAL avec son thésaurus. La passerelle ne peut être une bijection dans la mesure où le thésaurus ne comporte que des noms, et seulement les noms du domaine. En fait, seule une sous-partie du dictionnaire général pourra avoir un correspondant dans TMF. La correspondance pourra se faire entre « sens » et « TermSection ».

[à creuser]

11- Retroconversion d'OLIF-2

Notre modèle est englobant par rapport à OLIF-2.

Il est possible d'exprimer OLIF-2 dans notre modèle via une feuille de style XSLT.

[à faire : écrire la feuille de style pour le prouver]

12- DTD XML du modèle

[à faire quand le diagramme sera stabilisé]

13- Schéma XML du modèle

Le schéma XML est plus précis que la DTD dans la mesure où l'on peut typer les liens et ainsi rendre compte plus fidèlement du modèle conceptuel. En revanche, il est plus verbeux et le nombre de logiciels capable d'en faire usage est très limité.

[à faire quand le diagramme sera stabilisé, voir si on lui préfère un descriptif Relax-NG]

14- Références bibliographiques d'articles

Antoni-Lay MH., Francopoulo G., Zaysser L. 1994
A generic model for reusable lexicons : The Genelex project. Literary and Linguistic Computing.

Boguslavsky I. 2002
Some lexical issues of UNL. LREC.

Boitet C. 2002
The translation of examples, citations, definitions and glosses in the Papillon project. Journées Papillon. Tokyo.

Calzolari N., Fillmore C., Grishman R., Ide N., Lenci A., MacLeod C., Zampolli A. 2002
Towards best practice for Multiword Expressions in Computational Lexicons. LREC.

Dérouin MJ., Le Meur A. 2002.
Report on the revision of the lexicographical Standard ISO 1951. LREC.

Fellbaum C. 1998
WordNet : an electronic lexical database. MIT Press Cambridge, Mass

Fradin B. 2003
Nouvelles approches en morphologie. PUF. Paris.
Gaudin L., Guespin F. 2000
Initiation à la lexicologie française : de la néologie aux dictionnaires. Duculot. Bruxelles.
Genelex (consortium)
Rapport sur la couche morphologique 1991, syntaxique 1993, sémantique 1994.
Mangeot M., Sérasset G. 2001
Papillon lexical databases project : monolingual dictionaries and interlingual links. NLPRS. Tokyo.
Mel'cuk & al. 1984, 1988, 1992
Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques. Volumes 1, 2, 3. Presses de l'université de Montréal.
Mel'cuk I. 2000
Cours de morphologie générale. Volume-5. Presses de l'université de Montréal.
Mel'cuk I., Clas A., Polguère A. 1995
Introduction à la lexicologie explicative et combinatoire. Duculot. Bruxelles.
Peters C., Braschler M., Gonzalo J., Kluck M. 2002
Advances in Cross-Language Information Retrieval. Springer. Berlin.
Polguère A. 2000
Towards a theorically motivated general public dictionary of semantic derivation and collocation for French. Euralex 2000.
Pruvost J. 2000
Dictionnaires et nouvelles technologies. PUF. Paris.
Rey-Debove J. 1971
Etude linguistique et sémiotique des dictionnaires français contemporains. Mouton. The Hague.
Romary L. 2001
Towards an Abstract Representation of Terminological Data Collections – the TMF model. TAMA. Antwerp.
Saussure (de) F. 1974
Cours de linguistique générale : édition critique de Tullio de Mauro. Payot. Paris.
Silberztein M. 1993
Dictionnaires électroniques et analyse automatique de textes. Masson. Paris.

15- Références de sites comportant de multiples articles

EAGLES	www.ilc.cnr.it/EAGLES96/home.html
EDR/CRL	www2.crl.go.jp/kk/e416/EDR
Papillon	www.papillon-dictionary.org
Relax NG	www.oasis-open.org/committees/tc_home.php?wg_abbrev=relax-ng
Relex	www-igm.univ-mlv.fr/~unitex/linguistic_data.html
TEI	www.tei-c.org/P4X
UNL	www.unl.ias.unu.edu
WordNet	www.globalwordnet.org

16- Conclusion

Nous n'avons pas réinventé la roue. Nous avons pris les meilleures idées parmi les modèles que nous connaissions. Ces mécanismes descriptifs ont fait leurs preuves car ils sont mis en pratique depuis plusieurs années dans des équipes lexicographiques opérationnelles.
Chaque fois que c'était possible, nous avons utilisé les normes existantes comme les registres de catégories de données ISO 12620 ou Unicode.
Mais surtout, nous avons résisté à la tentation de la complexité inutile. Et, en effet, tout en étant puissant, le modèle reste simple.