

Distribution des Systèmes Lexicaux

Version 1.0

Sandrine Ollinger

sandrine.ollinger@atilf.fr

Alain Polguère

alain.polguere@univ-lorraine.fr

4 septembre 2017

Ressource distribuée sous licence :
Creative Commons – Attribution 4.0 International (CC BY 4.0)



Table des matières

Introduction	2
1 Organisation générale des fichiers d'export	2
2 Nœuds lexicaux du graphe	3
3 « Entrées » du Système Lexical	4
4 Structure polysémique des vocables	4
5 Caractéristiques grammaticales	6
6 Signifiants lexicaux	8
7 Étiquettes sémantiques	10
8 Formes propositionnelles	13
9 Liens de fonctions lexicales (FL)	14
10 Exemples lexicographiques	18
Références	20

Introduction

Cette documentation décrit le contenu des fichiers de distribution des *Systèmes Lexicaux*, les réseaux lexicaux développés au laboratoire ATILF. La lecture de ce document et l'exploitation des données lexicales qu'il présente présuppose une familiarisation minimale avec l'approche des Systèmes Lexicaux. Pour cela, nous recommandons d'effectuer, dans l'ordre, les lectures ci-dessous.

1. Lexicographie des Systèmes Lexicaux : Polguère (2014).
2. Projet RELIEF, dans le cadre duquel a débuté le programme de construction des Systèmes Lexicaux : Lux-Pogodalla et Polguère (2011).
3. Éditeur lexicographique Dicet, utilisé pour construire les bases lexicales SQL dont sont extraites les données présentées ici : Gader et al. (2012).
4. Structure de graphe petit-monde des Systèmes Lexicaux : Gader et al. (2014b).

Pour tout ce qui relève de la terminologie linguistique employée dans le présent document, on peut se reporter à Polguère (2016) – cf. *Index des notions*, pp. 371–379 de l'ouvrage.

Conventions d'écritures. Les notions linguistiques importantes apparaissent *en italique sans empattement* la première fois qu'elles sont utilisées. De façon plus générale, nous adoptons ici les conventions d'écriture de Polguère (2016).

1 Organisation générale des fichiers d'export

La version courante de l'export¹ comporte les 10 fichiers `.csv` et les 6 fichiers `.xml` ci-dessous, tous encodés en UTF-8.

- **Nœuds lexicaux du graphe**
01-lsnodes.csv
- **Entrées du Système Lexical**
02-lsentries.csv
- **Structure polysémique des vocables**
03-lscopolysemy-model.xml
04-lscopolysemy-rel.csv
- **Caractéristiques grammaticales**
05-lsgramcharac-model.xml
06-lsgramcharac-rel.csv
- **Signifiants lexicaux**
07-lswordform-model.xml
08-lswordforms.csv
- **Étiquettes sémantiques**
09-lssemlabel-model.xml
10-lssemlabel-rel.csv
- **Formes propositionnelles**
11-lspropform-rel.csv
- **Liens de fonctions lexicales**
12-lslf-model.xml
13-lslf-rel.csv
- **Exemples lexicographiques**
14-lsexsource-model.xml
15-lsex.csv
16-lsex-rel.csv

Les fichiers `.xml` contiennent le modèle **lexicologique** « théorique » (hiérarchie des caractéristiques grammaticales, système des fonctions lexicales, etc.). La structure de ces fichiers est précisée dans une DTD interne au début de chacun d'entre eux.

Les fichiers `.csv` contiennent les données du modèle **lexicographique** lui-même, qui décrivent le lexique de la langue (unités lexicales, leurs propriétés individuelles, les liens qui les

1. L'export est réalisé en PHP à l'aide de trois scripts programmés par S. Ollinger : `mainExport.php`, `DBAccessExport.php` et `GenerateExport.php`.

unissent, etc.). Leurs champs sont séparés par des tabulations et entourés de guillemets dactylographiques doubles ("...").

Le format d'export des Systèmes Lexicaux est le même, quelle que soit la langue modélisée. Le présent document (en français) se fonde sur les données du *Réseau Lexical du Français*, désormais *RL-fr*.

2 Nœuds lexicaux du graphe

Ce segment de l'export contient la définition de l'ensemble des nœuds du graphe. Ce sont par défaut des *lexies*, = *unités lexicales*, de la langue. Il peut cependant aussi s'agir de *clichés linguistiques*, de *collocations* ou même, marginalement, d'*expressions libres* (Mel'čuk, 2012). Cette information est encodée dans un unique fichier.

Fichier 01-lsnodes.csv

id	entry	lexnum	status	%
26162	26164	I.1	2	100
26163	26165	I.1	2	100
26164	26166	I.1	2	100

TAB. 1 : Extrait du fichier 01-lsnodes.csv

id	integer	ID du nœud lexical (typiquement, une lexie).
entry	integer	ID de l'entrée du Système Lexical correspondante (typiquement, un vocable – voir section 3 plus bas).
lexnum	vchar	Numéro lexicographique du nœud lexical.
status	integer	Statut d'avancement du travail lexicographique.
%	integer	Degré de confiance de l'information.

TAB. 2 : Description des données du fichier 01-lsnodes.csv

Le paramètre **status** prend une des quatre valeurs numériques ci-dessous :

- "3" – présence d'information lexicographique minimale ;
- "2" – description lexicographique véritablement débutée ;
- "1" – description lexicographique proche de la finalisation ;
- "0" – description lexicographique finalisée.

Vérification effectuée à chaque mise à jour

- Présence de lexies ayant un numéro lexicographique (**lexnum**) de valeur "0". Il s'agit de nœuds du graphe signalés par les lexicographes comme devant être détruits par la personne en charge de l'administration du modèle.

3 « Entrées » du Système Lexical

Un unique fichier, `02-lsentries.csv`, contient la spécification des « entrées » du Système Lexical, qui sont, dans la majorité des cas, des *vocables* de la langue (comme dans un dictionnaire courant). Chaque nœud du graphe étant associé à une entrée, il existe bien entendu marginalement des entrées correspondant à des collocations, des syntagmes libres, etc., et non à des vocables.

Dans ce fichier, sont notamment stockés les *noms lexicographiques* des vocables. Ainsi, le Tableau 3 ci-dessous montre comment sont spécifiés les noms lexicographiques des vocables À, À PROPOS, ADULTE_{Adj} et S'ALLONGER¹ (qui possède un vocable homonyme S'ALLONGER²).

Fichier `02-lsentries.csv`

id	addtoname	name	subscript	superscript	status	%
26164		à			1	100
26165		à propos			2	100
...						
26215		adulte	Adj		2	100
...						
26273	s'	allonger		1	2	100

TAB. 3 : Extrait du fichier `02-lsentries.csv`

id	integer	ID de l'entrée (typiquement, un vocable)
addtoname	varchar	Chaîne de caractères qui débute le nom lexicographique de l'entrée et qui ne doit pas entrer en ligne de compte dans le classement alphabétique de la nomenclature du Système Lexical : cf. <i>se</i> et <i>s'</i> , pour les verbes pronominaux français.
name	varchar	Corps du nom lexicographique de l'entrée.
subscript	varchar	Indice qui s'ajoute au nom lexicographique de l'entrée pour le spécifier grammaticalement, par exemple pour distinguer un vocable adjectival d'un vocable nominal nommé par le même signifiant linguistique (PETIT _{Adj} ~ PETIT _N).
superscript	varchar	Exposant (normalement, numérique) qui s'ajoute au nom lexicographique de l'entrée pour distinguer des homonymes (AVOCAT ¹ ~ AVOCAT ²).
status	integer	Statut d'avancement du travail lexicographique. Prend les mêmes valeurs que le paramètre <code>status</code> du Tableau 1, décrites plus haut p. 3.
%	integer	Degré de confiance de l'information.

TAB. 4 : Description des données du fichier `02-lsentries.csv`

4 Structure polysémique des vocables

La *structure polysémique* des vocables est modélisée dans deux fichiers :

- `03-lscopolysemy-model.xml` contient la définition des différents types et sous-types de liens de *copolysémie* ;

- 04-lscopolysemy-rel.csv contient l'ensemble des liens de copolysémie identifiés dans le Système Lexical.

L'ensemble des liens de copolysémie unissant les acceptions d'un vocable donné constitue la structure polysémique de ce vocable. Notons que cette structure est hiérarchique, mais non nécessairement arborescente : il peut exister des cas (rares) de lexies qui sont subordonnées à plus d'une autre acception de leur vocable.

Fichier 03-lscopolysemy-model.xml

```
<model>
  <type id="3" name="causation" order="1" semantics="2" derivation="1"/>
  <type id="6" name="résultat" order="2" semantics="2" derivation="1"/>
  <type id="7" name="conversion" order="3" semantics="2" derivation="1"/>
  <type id="4" name="spécialisation" order="4" semantics="2" derivation="1">
    <subtype id="10" name="sous-sens"/>
  </type>
  ...
</model>
```

FIG. 1 : Extrait du fichier 03-lscopolysemy-model.xml

id	integer	ID du lien de copolysémie.
name	varchar	Nom du lien de copolysémie.
order	integer	Position absolue du lien de copolysémie dans une énumération des liens.
semantics	integer	Poids sémantique du lien de copolysémie : de "0" (poids nul) à "2" (poids maximal).
derivation	boolean	Pour un lien de copolysémie $L_1 \rightarrow L_2$, l'acception L_2 est/n'est pas considérée comme construite sémantiquement à partir de l'acception L_1 .

TAB. 5 : Description des données du fichier 03-lscopolysemy-model.xml

Fichier 04-lscopolysemy-rel.csv

source	target	type	subtype
44006	44007	1	4
32937	28866	5	
26291	30859	1	7

TAB. 6 : Extrait du fichier 04-lscopolysemy-rel.csv

source	integer	ID de la source du lien de copolysémie.
target	integer	ID de la cible du lien de copolysémie.
type	integer	ID du type du lien de copolysémie, tel que défini par <code>type</code> dans le modèle (voir plus haut, fichier 03-lscopolysemy-model.xml).
subtype	integer	ID du sous-type du lien de copolysémie, tel que défini par <code>subtype</code> dans le modèle.

TAB. 7 : Description des données du fichier 04-lscopolysemy-rel.csv

Le lien logique entre `type` et `subtype` n'est pour l'instant pas encodé de façon explicite dans la base lexicographique. Il n'est donc pas récupérable automatiquement². Comme l'éditeur lexicographique ne dispose pas de l'information nécessaire pour contrôler ce lien logique, il peut arriver que les lexicographes créent des associations « `<type>` : `<subtype>` » invalides – par exemple, « métaphore : ensemble de ».

Vérification effectuée à chaque mise à jour

- Y a-t-il de mauvaises associations entre types et sous-types ?

5 Caractéristiques grammaticales

Les caractéristiques grammaticales associées à chaque nœud du Système Lexical (marques d'usage, partie du discours, etc.) sont organisées au sein d'une hiérarchie. La caractérisation grammaticale est exportée dans deux fichiers :

- `05-lsgramcharac-model.xml` contient la hiérarchie des caractéristiques grammaticales disponibles pour l'étiquetage grammatical des nœuds lexicaux ;
- `06-lsgramcharac-rel.csv` contient l'ensemble des étiquetages grammaticaux (associations entre les nœuds lexicaux et leurs caractéristiques grammaticales).

Fichier `05-lsgramcharac-model.xml`

```
<model>
  <characteristic id="2" name="MARQUES D'USAGE" type="0" status="0">
    <characteristic id="3" name="Marques d'usage langagier" type="0" status="0">
      <characteristic id="6" name="spéc" type="1" status="0"/>
      ...
    </characteristic>
    ...
  </characteristic>
  ...
</model>
```

FIG. 2 : Extrait du fichier `05-lsgramcharac-model.xml`

2. Un bricolage jouant entre `position` et `id` se fait actuellement dans le script d'export afin d'associer le bon `subtype` au `type` correspondant dans la structure XML du fichier `03-lscopolysemy-model.xml`.

id	integer	ID de la caractéristique grammaticale.
name	vchar	Nom de la caractéristique grammaticale.
type	integer	Typage de la caractéristique grammaticale. Dans l'état courant du modèle, trois types sont utilisés : "0" pour 'autre caractéristique grammaticale' (valeur par défaut), "1" pour 'marque d'usage' et "2" pour 'partie du discours'.
status	integer	Statut lexicologique de la caractéristique grammaticale. Dans l'état courant du modèle, seulement deux valeurs sont utilisées : "0" = caractéristique grammaticale valide <i>vs</i> "1" = caractéristique grammaticale non encore validée.

TAB. 8 : Description des données du fichier `05-lsgramcharac-model.xml`

L'algorithme de création du fichier `05-lsgramcharac-model.xml` doit être amélioré. Pour l'instant, il n'est pas en mesure de prendre en charge la création de nouvelles caractéristiques grammaticales si celles-ci impliquent un niveau d'imbrication supérieur à 3 parents, qui est la limite actuelle (cas des parties du discours).

Fichier `06-lsgramcharac-rel.csv`

id	usagenote	POS	phraseolstruc	embededlex	othercharac
30750		20			(26,50)
30751		24			
30752		20			(28)
30753		20			(28)
30754		20			(28)
30755		20			(28)
30756		23			
30758		21	NC Prép NC	(46192,46193)	(26,166)

TAB. 9 : Extrait du fichier `06-lsgramcharac-rel.csv`

id	integer	ID du nœud lexical considéré.
usagenote	array of integers	Liste des ID des marques d'usage (spéc , fam , etc.) du nœud lexical.
POS	integer	ID de la partie du discours (<i>part of speech</i>).
phraseolostruc	vchar	Chaîne de caractères identifiant le patron de structure lexico-syntaxique, pour une locution.
embededlex	array of integers	Liste des ID des lexies contenues dans la structure lexico-syntaxique, pour une locution.
othercharac	array of integers	ID des autres caractéristiques grammaticales.

TAB. 10 : Description des données du fichier `06-lsgramcharac-rel.csv`

Comme on le voit, les structures lexico-syntaxiques des *locutions* (Pausé, 2017) sont intégrées dans le fichier `06-lsgramcharac-rel.csv`. Ainsi, la dernière ligne du Tableau 9, plus haut, indique que 「LEVÉE DE BOUCLIER」 (ID 30758) est une locution nominale (ID 21) dont la structure lexico-syntaxique NC Prép NC comprend les deux lexies LEVÉE IV (ID 46192) et BOUCLIER (ID 46193).

Pour l’instant, l’export ne spécifie pas à quel nœud lexical de la structure lexico-syntaxique de la locution correspond chacune des lexies enchâssées de la colonne `embeddlex`. Par exemple, dans l’exemple analysé ci-dessus, il n’est pas indiqué explicitement que LEVÉE **IV** est le premier nom commun (NC) de la chaîne NC **Prép** NC et BOUCLIER le second. De plus, la structure n’est pas encodée comme un arbre syntaxique véritable, tel que représenté dans la Figure 3 :

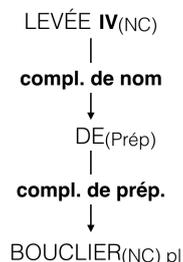


FIG. 3 : Structure lexico-syntaxique de la locution 「LEVÉE DE BOUCLIER」

Ce type d’information sera disponible dans une version ultérieure de l’export.

Vérifications effectuées à chaque mise à jour

- Y a-t-il des lexies associées à plusieurs parties du discours ?
- Y a-t-il des lexies sans partie du discours ?
- Y a-t-il des caractéristiques grammaticales du modèle non utilisées dans le Système Lexical ?

6 Signifiants lexicaux

La morphologie des lexies est implantée de façon assez sophistiquée dans les Systèmes Lexicaux du type RL-fr au moyen de modèles morphologiques, qui permettent notamment de générer toutes les variantes flexionnelles d’une lexie. La stratégie lexicographique adoptée pour modéliser la morphologie des lexies est décrite dans Gader et al. (2014a).

Les règles morphologiques du modèle ne sont pas présentes dans l’export, où seuls figurent les *mots-formes* associés à chaque lexie et les traits morphologiques utilisés pour caractériser les mots-formes. Cette information est répartie en deux fichiers :

- `07-lswordform-model.xml` contient l’ensemble des traits morphologiques (*feature*), regroupés en catégories (*category*) – il peut s’agir de grammèmes véritables (*Singulier*, *Pluriel*...) aussi bien que de paramètres d’usage (*graphie rectifiée*, *forme élidée*...);
- `08-lswordforms.csv` contient tous les mots-formes de chaque lexie, chaque mot-forme étant associé à un ensemble de traits du fichier `07-lswordform-model.xml`.

Fichier 07-lswordform-model.xml

```
<model>
<category id="1" name="Mode" pivot-name="mood">
<feature id="1" name="Indicatif" pivot-name="indicative" position="1"/>
<feature id="2" name="Conditionnel" pivot-name="conditional" position="2"/>
<feature id="3" name="Subjonctif" pivot-name="subjunctive" position="3"/>
<feature id="4" name="Impératif" pivot-name="imperative" position="4"/>
<feature id="5" name="Infinitif" pivot-name="infinitive" position="5"/>
<feature id="6" name="Participe" pivot-name="participle" position="6"/>
</category>
...
</model>
```

FIG. 4 : Extrait du fichier 07-lswordform-model.xml

id	integer	ID d'un trait (grammatical, formel...) ou d'une catégorie de traits.
name	vchar	Nom d'un trait ou d'une catégorie de traits formulé dans la langue modélisée par le Système Lexical.
pivot-name	vchar	Nom d'un trait ou d'une catégorie de traits formulé au moyen d'un terme anglais pouvant servir de pivot dans le contexte de la mise en relation de plusieurs Systèmes Lexicaux (modélisant des langues différentes).
position	integer	Position absolue d'un trait dans une énumération de traits (cf. paramètre <code>features</code> du fichier 08-lswordforms.csv ci-dessous).

TAB. 11 : Description des données du fichier 07-lswordform-model.xml

Fichier 08-lswordforms.csv

id	features	signifier
26164	(1,10,15,16)	abandonna
26164	(1,10,13,16)	abandonnai
26164	(1,8,15,17)	abandonnaient

TAB. 12 : Extrait du fichier 08-lswordforms.csv

id	integer	ID du mot-forme.
features	array of integers	Liste ordonnée de façon normalisée des ID des traits (grammaticaux, formels...) portés par le mot-forme.
signifier	vchar	Signifiant du mot-forme.

TAB. 13 : Description des données du fichier 08-lswordforms.csv

Vérification effectuée à chaque mise à jour

- Présence de mots-formes associés à des règles morphologiques n'existant pas dans le modèle.

7 Étiquettes sémantiques

Les Systèmes Lexicaux utilisent une méthode de classification sémantique des lexies fondée sur une hiérarchie d'*étiquettes sémantiques*. Plus précisément, chaque étiquette est une instance d'une classe d'étiquettes, les classes étant organisées au sein d'une hiérarchie à héritage multiple. Ce système de classification sémantique est décrit dans Polguère (2011)³. Il réside dans deux fichiers de l'export :

- 09-lssemlabel-model.xml contient la hiérarchie de classes d'étiquettes sémantiques avec leurs instances (c'est-à-dire, les étiquettes elles-mêmes) ;
- 10-lssemlabel-rel.csv contient l'étiquetage des lexies individuelles.

Fichier 09-lssemlabel-model.xml

```
<model>
  <class id="534" name="QQCH." status="1" semfield="0" inheritancetype="0" comment="">
    <class id="253" name="ENTITÉ" status="1" semfield="0" inheritancetype="0" comment="">
      ...
      <instance id="281" name="entité" status="1" derivation="---" acttype="" comment=""/>
    </class>
    <class id="273" name="FAIT" status="1" semfield="0" inheritancetype="0" comment="">
      <class id="35" name="ACTION" status="1" semfield="0" inheritancetype="0"
        comment="Fait ponctuel dont le premier actant est agent -> X est causateur volontaire">
        <class id="27" name="ACTE_DE_COMMUNICATION" status="1" semfield="1" inheritancetype="0" comment="">
          <class id="159" name="COMMUNICATION_LANGAGIÈRE" status="1" semfield="1" inheritancetype="0" comment="">
            ...
            <instance id="177" name="communication langagière" status="1" derivation="---" acttype="0" comment=""/>
            <instance id="178" name="dire [qqch.]" status="1" derivation="V0" acttype="0" comment=""/>
            <instance id="179" name="qui exprime un énoncé d'un certain type" status="1" derivation="A1" acttype="0" comment=""/>
            <instance id="974" name="relatif à une communication langagière" status="1" derivation="A0" acttype="0" comment=""/>
          </class>
          ...
          <instance id="27" name="acte de communication" status="1" derivation="---" acttype="0" comment=""/>
          <instance id="885" name="communiquer [qqch.]" status="1" derivation="V0" acttype="" comment=""/>
        </class>
        ...
        <instance id="37" name="action" status="1" derivation="---" acttype="1" comment=""/>
        <instance id="926" name="faire qqch." status="1" derivation="V0" acttype="0" comment=""/>
      </class>
      ...
      <instance id="303" name="fait" status="1" derivation="---" acttype="" comment=""/>
      <instance id="793" name="avoir lieu" status="1" derivation="V0" acttype="0" comment="Étiquette notamment de verbes supports de type Func0."/>
    </class>
    <instance id="595" name="qqch." status="1" derivation="---" acttype="1" comment=""/>
  </class>
</model>
```

FIG. 5 : Extrait du fichier 09-lssemlabel-model.xml

3. Voir aussi Milićević et Polguère (2010) pour ce qui est de l'héritage multiple comme encodage de l'*ambivalence sémantique* de certaines lexies.

class	id	integer	ID de la classe d'étiquettes sémantiques.
	name	varchar	Nom de la classe.
	status	integer	Statut lexicologique de la classe. Dans l'état courant du modèle, seulement deux valeurs sont utilisées : "0" = classe valide <i>vs</i> "1" = classe non encore validée.
	semfield	boolean	Indique si la classe est utilisable pour identifier une champ sémantique du modèle.
	inheritancetype	integer	Type de connexion à la classe-mère considérée. Dans l'état courant du modèle, trois valeurs sont utilisées : "0" = héritage simple (une seule mère), "1" = héritage multiple inclusif (<i>and/or</i>), "2" = héritage multiple exclusif (<i>or</i>).
	comment	mediumtext	Commentaires éventuels sur la classe.
instance	id	integer	ID de l'instance de classe = de l'étiquette.
	name	varchar	Nom de l'étiquette.
	status	integer	Statut lexicologique de l'étiquette – mêmes valeurs que pour les statuts de classes plus haut.
	derivation	varchar	Nature sémantico-grammaticale de l'étiquette, exprimée en termes de fonctions lexicales – voir explications plus bas.
	acttype	boolean	Indique si l'étiquette peut être utilisée pour typer une variable actancielle dans une définition lexicographique.
	comment	mediumtext	Commentaires éventuels sur l'étiquette.

TAB. 14 : Description des données du fichier `09-lssemlabel-model.xml`

Tel qu'expliqué en détail dans Polguère (2011), une étiquette sémantique type est nominale et est utilisable uniquement pour étiqueter des lexies nominales. Chaque classe d'étiquette, cependant, regroupe potentiellement une grappe d'étiquettes sémantiques qui permettent d'étiqueter des dérivés sémantiques (verbaux, adjectivaux... et même nominaux) des lexies étiquetées par l'étiquette nominale « de base » de la classe. Le paramètre `derivation` associé à chaque étiquette spécifie son statut (de base ou dérivational) dans la classe. Le Tableau 15 ci-dessous donne les valeurs possibles de ce paramètre, dans l'état courant du modèle lexicologique.

---	Étiquette de base de la classe d'étiquettes pour des lexies nominales L.
Convij	Conversifs de L.
VOConvij	Verbalisations conversives de L.
VO	Verbalisations de L.
A0	Adjectivations de L.
Adv0	Adverbalisations de L.
S1	Noms de 1 ^{er} actant de L.
S2	Noms de 2 ^e actant de L.
S3	Noms de 3 ^e actant de L.
A1	Adjectifs qui dénotent le 1 ^{er} actant de L.
A2	Adjectifs qui dénotent le 2 ^e actant de L.
A3	Adjectifs qui dénotent le 3 ^e actant de L.
Adv1	Adverbes qui signifient '[faire qqch.] en L-ant'.
A2Manif	Adjectifs qui dénotent N dans <i>L se manifeste dans N</i> .
Able1	Adjectifs qui dénotent un 1 ^{er} actant possible de L.
Able2	Adjectifs qui dénotent un 2 ^e actant possible de L.
Caus	Verbes qui signifient 'causer L'.
ElementOf	Noms qui dénotent un élément de L.
SetOf	Noms qui dénotent un ensemble de L.
PartOf	Noms qui dénotent une partie de L.
Small	Noms qui dénotent un petit L.
Type	Noms qui dénotent un type de L.
Claus	Énoncé (= clausatif) « L ».

TAB. 15 : Valeurs possibles du paramètre *derivation*

La structure XML du fichier *09-lssemblabel-model.xml* encode une hiérarchie de classes d'étiquettes sémantiques dont un échantillon est visualisé dans la Figure 6 ci-dessous⁴.

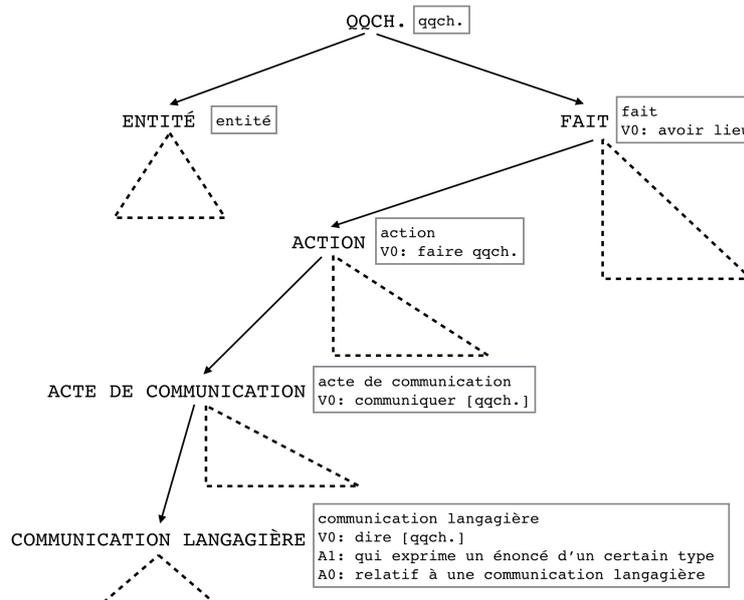


FIG. 6 : Échantillon de la hiérarchie des classes d'étiquettes sémantiques

4. Les instances de chaque classe d'étiquettes apparaissent dans une boîte à côté du nom de la classe.

Fichier 10-lssemlabel-rel.csv

sense	label	%
26162	152	60
26164	734	60
26166	699	100
26167	583	60

TAB. 16 : Extrait du fichier 10-lssemlabel-rel.csv

sense	integer	ID de la lexie étiquetée.
label	integer	ID de l'étiquette de la lexie.
%	integer	Degré de confiance de l'information.

TAB. 17 : Description des données du fichier 10-lssemlabel-rel.csv

Vérifications effectuées à chaque mise à jour

- Y a-t-il des étiquettes non utilisées ?
- Y a-t-il des guillemets ou apostrophes droits dans les commentaires ?

8 Formes propositionnelles

Les *formes propositionnelles* associées à chaque lexie sont entrées avec l'éditeur lexicographique sous le contrôle d'un modèle hiérarchique des patrons de formes propositionnelles admissibles. Cela permet d'assurer une certaine systématique dans la description des *structures actancielles* lexicales dans l'ensemble du Système Lexical.

L'export ne contient pas le modèle de patrons de formes propositionnelles, mais un unique fichier – 11-lspropform-rel.csv – qui rassemble l'ensemble de toutes les associations « lexie-forme propositionnelle ».

Fichier 11-lspropform-rel.csv

id	propform	tildevalue	%	actantslist
26162	[\$1] ~ \$2		100	(\$1=X,\$2=Y)
26163	[\$1] ~ de \$2		100	(\$1=X,\$2=Y)
26166	~		100	()
26167	\$1 ~ \$2	abîme	100	(\$1=X,\$2=Y)
26168	\$1 ~ \$2 à \$3	abonne	100	(\$1=X,\$2=Y,\$3=Z)

TAB. 18 : Extrait du fichier 11-lspropform-rel.csv

<code>id</code>	integer	ID du nœud lexical considéré (la lexie vedette d'un article lexicographique).
<code>propform</code>	varchar	Forme propositionnelle proprement dite, où le tilde (~) indique la position de la lexie vedette dans la formule et \$1, \$2... indiquent les positions actancielles pour le 1 ^{er} , 2 ^e ... actant.
<code>tildevalue</code>	varchar	Optionnel : forme devant être utilisée à la place du nom lexicographique de la lexie vedette pour instancier le tilde de la formule dans la présentation « textuelle » de l'information lexicale, au sein d'un article lexicographique.
<code>%</code>	integer	Degré de confiance de l'information.
<code>actantslist</code>	array of varchars	Liste des variables X, Y... devant être utilisées à la place de \$1, \$2... dans la présentation de l'information lexicale au sein d'un article lexicographique.

TAB. 19 : Description des données du fichier `11-lspropform-rel.csv`

Ainsi, selon les explications données ci-dessus, la version article lexicographique de la dernière ligne du Tableau 18 se formulera de la façon suivante : **X abonne Y à Z.**

Vérification effectuée à chaque mise à jour

- Vérifier la validité des espaces situées dans les formules de formes propositionnelles.

9 Liens de fonctions lexicales (FL)

Pour une introduction succincte à la notion de *fonction lexicale* – désormais *FL* –, on pourra se reporter à Polguère (2016), pp.195–210. Les liens de FL tissent l'ossature des Systèmes Lexicaux. Ils résident dans deux fichiers de l'export :

- `12-lslf-model.xml` contient le modèle hiérarchique des FL : chaque FL appartient à une « famille » et chaque famille est elle-même élément d'un « groupe » de familles ;
- `13-lslf-rel.csv` contient l'ensemble de liens de FL entre lexies individuelles.

Fichier 12-lslf-model.xml

```
<model>
  ...
  <group>
    <family id="3" name="Syn">
      <lexicalfunction id="2" name="Syn" status="0" linktype="paradigmatic"
        standardness="simple standard" semantics="2">
        <![CDATA[<font size=3><b>Syn</b>]]>
      </lexicalfunction>
      ...
    </family>
    ...
  </group>
  ...
  <group>
    <family id="15" name="S1">
      <lexicalfunction id="31" name="S_1" status="0" linktype="paradigmatic"
        standardness="simple standard" semantics="2">
        <![CDATA[<font size=3><b>S<sub>1</sub></b>]]>
      </lexicalfunction>
      ...
    </family>
    ...
  </group>
  ...
</model>
```

FIG. 7 : Extrait du fichier 12-lslf-model.xml

id	integer	ID de la fonction lexicale ou de la famille de fonctions lexicales.
name	varchar	Nom de la fonction lexicale ou de la famille de fonctions lexicales. Les caractères « _ » et « ^ » introduisent, respectivement, un indice <i>vs</i> un exposant – par exemple, S_1 pour la fonction lexicale S₁ .
status	integer	Statut lexicologique de la fonction lexicale. Dans l'état courant du modèle, seulement deux valeurs sont utilisées : "0" = fonction lexicale valide <i>vs</i> "1" = fonction lexicale non encore validée.
linktype	varchar	Typage de la fonction lexicale relativement à la nature sémantique ou combinatoire du lien qu'elle tisse ; dans l'état courant du modèle, seulement deux valeurs sont utilisées : "paradigmatic" <i>vs</i> "syntagmatic".
standardness	varchar	Typage de la fonction lexicale relativement à son caractère formel : simple ou complexe, plus ou moins standard, etc.
semantics	integer	Poids sémantique des liens lexicaux tissés par la fonction lexicale : de "0" (poids nul) à "2" (poids maximal).
CDATA	html	Variante formatée du nom de fonction lexicale (paramètre name ci-dessus), pour affichage dans un article lexicographique.

TAB. 20 : Description des données du fichier `12-lslf-model.xml`

Deux remarques sur l'export dans `12-lslf-model.xml` :

1. les éventuelles familles de FL de la base du Système Lexical qui ne regroupent aucune FL ne sont pas intégrées à l'export ;
2. dans les noms de FL, les occurrences du caractère « & » dans la base du Système Lexical ont été remplacées par des « & ».

Rappelons qu'une fonction lexicale donnée **f** s'applique aux lexies L de la langue – **f(L)** – pour retourner une valeur {V} qui est un ensemble de lexies (ou expressions linguistiques) exprimant le contenu sémantico-syntaxique de **f** relativement à L. Cela se représente par la formule suivante, qui est une métaphore formelle de l'application d'une fonction mathématique :

$$\mathbf{f}(L) = \{V\}.$$

Pour une illustration, voir la formule (1) plus bas, p. 17.

Fichier 13-lslf-rel.csv

source	lf	target	form	separator	merged	syntacticframe	constraint	position
35193	5	28895		,	0			1
35193	5	35202		,	0			2
35193	524	26369		,	0			1
35193	524	26520		,	0			2
35193	524	54368		;	0			3
35193	524	26980		;	0			4
35193	524	28171		;	0			5
35193	74	30608		,	0	de ~		1
35193	74	33239		,	0	de ~		2
35193	74	30335		;	0	de ~		3
35193	228	28235		,	0	ART ~		1
35193	228	47505		,	0	ART ~		2

TAB. 21 : Extrait du fichier 13-lslf-rel.csv

source	integer	ID de la lexie source du lien de FL.
lf	integer	ID de la FL.
target	integer	ID de la cible (généralement une lexie) du lien de FL.
form	vchar	Optionnel : formulation qui devrait être substituée au nom lexicographique de la cible du lien dans un article lexicographique.
separator	vchar	Séparateur entre cette cible et l'éventuelle cible précédente dans la liste des éléments de la valeur de l'application $\mathbf{f}(L)$. Les séparateurs possibles sont : « , » (faible distance sémantique entre cette cible et la précédente), « ; » (distance sémantique assez notable) et « < » (distance sémantique correspondant à une augmentation d'amplitude, intensité...).
merged	boolean	Booléen qui indique s'il s'agit d'une valeur non fusionnée (« 0 ») – cas par défaut – ou fusionnée (« 1 »), dans le cas d'une FL syntagmatique.
syntacticframe	vchar	Optionnel : régime syntaxique de la cible du lien de FL.
constraint	vchar	Optionnel : contraintes grammaticales, stylistiques, etc.
position	integer	Position relative de la cible dans la liste des cibles pour la même application $\mathbf{f}(L)$.

TAB. 22 : Description des données du fichier 13-lslf-rel.csv

Les liens de FL connectent des lexies individuelles, mais dans les faits, toute application de FL \mathbf{f} à une lexie donnée L – c'est-à-dire $\mathbf{f}(L)$ – retourne une valeur qui est un **ensemble** de lexies. Il est donc nécessaire, pour chaque lien, d'indiquer comment la cible se positionne à l'intérieur de la liste des éléments de la valeur de $\mathbf{f}(L)$ – cf. valeur de **position**. Par exemple, les lignes 9 à 11 du Tableau 21 ci-dessus se réaliseront dans un article lexicographique de la façon suivante :

- (1) **Sing**(*soupe*) = *assiette*¹ **I.b** [*de -*], *assiettée* [*de -*]; *bol*¹ **I.b** [*de -*].

Vérification effectuée à chaque mise à jour

- Y a-t-il des FL du modèle non utilisées ?

10 Exemples lexicographiques

Les exemples lexicographiques sont introduits comme des entités informationnelles autonomes dans les Systèmes Lexicaux. L'exemplification des lexies se fait par association des lexies individuelles à un ou plusieurs exemples – voir Lux-Pogodalla (2014).

L'information sur l'exemplification réside dans trois fichiers de l'export :

- 14-lsexsource-model.xml contient le modèle des sources des exemples lexicographiques (notamment, les corpus textuels) ;
- 15-lsex.csv contient les exemples lexicographiques proprement dits ;
- 16-lsex-rel.csv contient les associations lexie → exemple.

Fichier 14-lsexsource-model.xml

```
<model>
...
<section id="3" name="Citations de corpus journalistiques hors Frantext">
  <source id="3" name="L'Est Républicain"/>
  <source id="4" name="Chambers-Rostand"/>
</section>
...
</model>
```

FIG. 8 : Extrait du fichier 14-lsexsource-model.xml

id	integer	ID de corpus ou de section dans l'ensemble des corpus (= type de corpus).
name	varchar	Nom du corpus ou de la section.

TAB. 23 : Description des données du fichier 14-lsexsource-model.xml

Fichier 15-lsex.csv

id	source	status	content	title	authors	location	date
13	1	1	<html><body><p>Nous voulions kidnapper le gérant et le juger. Nous avons entamé une campagne de propagande contre <<le gérant fasciste de Seyssinet>>.</p></body></html>	Voyage au bout de la révolution : de Pékin à Sochaux	Brière-Blanchet, Claire	p. 116	2009
16	2	1	<html><body><p>Les ravisseurs ont profité d'une absence de vigilance de la mère pour kidnapper la petite Amanda.</p></body></html>	...		http://www.yozone.fr/spip.php?article4460	02/2008
135	5	1	<html><body><p>Quoi de mieux que de finir la semaine par un petit basket entre midi et 2 ? ! Alors venez mettre quelques paniers, avec de la technique, du style, ou simplement de la chance !</p></body></html>	...		http://www.savoie-technolac.com/acs/10/127-ac-culture-sport.htm	16/10/2012

TAB. 24 : Extrait du fichier 15-lsex.csv

id	integer	ID de l'exemple.
source	integer	ID de la source.
status	integer	Statut lexicographique de l'exemple. Dans l'état courant du modèle, seulement deux valeurs sont utilisées : "0" = exemple valide <i>vs</i> "1" = exemple non encore validé.
content	html	Contenu textuel de l'exemple. Celui-ci est stocké en HTML, car il peut contenir du formatage (italiques, gras, mises en exposant, etc.).
title	varchar	Titre de l'ouvrage (ou assimilé) d'où provient l'exemple.
authors	varchar	Auteur(s) de l'exemple.
location	varchar	Localisation de la publication (éditeur, page web, etc.).
date	date	Date associée à chaque exemple. La spécification de la date est scindée en trois champs dans la base lexicographique du Système Lexical ; elle occupe ici un seul champ, sous la forme AAAA, MM/AAAA ou JJ/MM/AAAA.

TAB. 25 : Description des données du fichier 15-lsex.csv

- Seuls les exemples associés à au moins un nœud du graphe ont été exportés.

Fichier 16-lsex-rel.csv

id	exemple	occurrence	position	%
44432	13882	163,171 ;	0	100
44432	13746	66,75 ;	1	100
44432	13779	87,95 ;	2	100
44432	13804	16,24 ;	3	100
44432	13859	302,310 ;	4	100
44433	13750	120,128 ;	0	100
44433	13821	197,206 ; 273,282 ;	1	100

TAB. 26 : Extrait du fichier 16-lsex-rel.csv

id	integer	ID de la lexie exemplifiée.
exemple	integer	ID d'un des exemples illustratifs qui est associé à la lexie.
occurrence	varchar	Segment(s) de texte de l'exemple où apparaît la lexie. Chaque segment est indiqué par la position dans l'exemple du premier caractère du segment, suivie du caractère « , », lui-même suivi de la position du dernier caractère. Les segments sont séparés par le caractère « ; ».
position	integer	Position relative de l'exemple, parmi les autres exemples, dans une visualisation de type article lexicographique.
%	integer	Degré de confiance de l'information.

TAB. 27 : Description des données du fichier 16-lsex-rel.csv

Le paramètre **occurrence** contient plus d'un segment dans un des deux cas suivants :

- l'exemple contient plusieurs occurrences de la lexie qu'il sert à illustrer – par exemple, *Être ou ne pas être*, dans un exemple illustrant le lexème verbal ÊTRE_V I ;

- la lexie est exprimée sous la forme d’un syntagme disjoint – par exemple, *avait maladroitement renversé*, dans un exemple illustrant le lexème RENVERSER 1.1, ou *jouer souvent des coudes*, dans un exemple illustrant la locution 「JOUER DES COUDES」.

Vérification effectuée à chaque mise à jour

- Présence de guillemets droits dans les champs `text`, `title`, `authors` et `location` de la table `ln_ex` de la base lexicographique source de l’export.

Références

- Nabil Gader, Aurore Koehl et Alain Polguère. A Lexical Network with a Morphological Model in It. Dans *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex IV)*, pages 154–165, Dublin, 2014a. Association for Computational Linguistics and Dublin City University.
- Nabil Gader, Veronika Lux-Pogodalla et Alain Polguère. Hand-Crafting a Lexical Network With a Knowledge-Based Graph Editor. Dans *Proceedings of the Third Workshop on Cognitive Aspects of the Lexicon (CogALex III)*, pages 109–125, Mumbai, 2012. The COLING 2012 Organizing Committee.
- Nabil Gader, Sandrine Ollinger et Alain Polguère. One Lexicon, Two Structures : So What Gives ? Dans Heili Orav, Christiane D. Fellbaum et Piek Vossen, dir., *Proceedings of the Seventh Global Wordnet Conference (GWC2014)*, pages 163–171, Tartu (Estonie), 2014b. Global WordNet Association.
- Veronika Lux-Pogodalla. Intégration relationnelle des exemples lexicographiques dans un réseau lexical. Dans Brigitte Bigi, dir., *Actes de TALN 2014*, pages 586–591, Marseille, 2014. Laboratoire Parole et Langage, Aix-en-Provence.
- Veronika Lux-Pogodalla et Alain Polguère. Construction of a French Lexical Network : Methodological Issues. Dans *Proceedings of the First International Workshop on Lexical Resources, WoLeR 2011. An ESSLLI 2011 Workshop*, pages 54–61, Ljubljana, Slovenia, 2011.
- Igor Mel’čuk. Phraseology in the Language, in the Dictionary, and in the Computer. Dans Koenraad Kuiper, dir., *Yearbook of Phraseology*, volume 3, pages 31–56. De Gruyter Mouton, Berlin, 2012.
- Jasmina Milićević et Alain Polguère. Ambivalence sémantique des noms de communication langagière du français. Dans F. Neveu, V. Muni Toke, J. Durand, T. Klingler, L. Mondada et S. Prévost, dir., *Actes de la section « Lexique et morphologie » du 2^e Congrès Mondial de Linguistique Française (CMLF’10)*, pages 1029–1050, La Nouvelle-Orléans, 12–15 juillet 2010. Institut de Linguistique Française (ILF).
- Marie-Sophie Pausé. *Structure lexico-syntaxique des locutions du français et incidence sur leur combinatoire*. Thèse de doctorat, Université de Lorraine, Nancy, 2017.
- Alain Polguère. Classification sémantique des lexies fondée sur le paraphrasage. *Cahiers de lexicologie*, 98 :197–211, 2011.
- Alain Polguère. From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*, 27(4) :396–418, 2014.
- Alain Polguère. *Lexicologie et sémantique lexicale. Notions fondamentales*. Paramètres. Les Presses de l’Université de Montréal, Montréal, 3^e édition, 2016.