

## Modélisation de dictionnaires spécialisés à l'aide de la lexicologie explicative et combinatoire (LEC)

Marie-Claude L'Homme

coll. Iveth Carreño, Jeanne Dancette, Philippe Hanscom, Anne-Laure Jousse, Chantal Lemay

Observatoire de Linguistique Sens <-> Texte

<http://www.olst.umontreal.ca>

(OLST) Équipe de terminologie (ÉCLECTIQ)

Département de linguistique et de traduction

Université de Montréal

## Plan

- **Expérience 1** : conversion d'un dictionnaire papier existant (*Dictionnaire bilingue de la distribution*, Dancette et Réthoré 2000)
  - Description du dictionnaire papier
  - Modèle relationnel
  - Modélisation des relations sémantiques au moyen des fonctions lexicales (Mel'cuk et al. 1988; 1995).
  - Bilan
- **Expérience 2** : Élaboration d'un dictionnaire d'informatique
  - Objectifs du dictionnaire
  - Sélection des entrées
  - Distinctions sémantiques
  - Relations sémantiques entre termes

## Plan

- **Expérience 1** : conversion d'un dictionnaire papier existant (*Dictionnaire bilingue de la distribution*, Dancette et Réthoré 2000)
  - Description du dictionnaire papier
  - Modèle relationnel
  - Modélisation des relations sémantiques au moyen des fonctions lexicales (Mel'cuk et al. 1988; 1995).
  - Bilan
- **Expérience 2** : Élaboration d'un dictionnaire d'informatique
  - Objectifs du dictionnaire
  - Sélection des entrées
  - Distinctions sémantiques
  - Relations sémantiques entre termes

## Dictionnaire bilingue de la distribution (description)

- Dancette et Réthoré (2000). *Dictionnaire bilingue de la distribution*, Montréal : Les Presses de l'Université de Montréal.
- Entrées anglaises – entrées françaises
- Le corps des articles en français
- 350 articles
- 3500 termes à l'intérieur des articles
- Destiné aux traducteurs, professionnels et professeurs

## Dictionnaire bilingue de la distribution (article type) (1)

&1 **WHITE GOODS**

2 **PRODUITS BLANCS**

3 **Définition** : Ensemble des gros appareils électroménagers destinés plus particulièrement à la salle de lavage ou à la cuisine.

4 **Précisions sémantiques** : Ces appareils sont appelés **produits blancs** en raison de la couleur dans laquelle ils sont le plus souvent fabriqués.

5 **Relations internationales** :

Les **produits blancs** sont classés dans la catégorie des **biens durables** (durable goods), qui s'opposent aux **biens non durables** (non-durable goods) et s'en distinguent par leur durée de vie. Voir **PRODUCT**, **GOOD**.

L'expression **produits bruns** (brown goods) désigne le groupe d'appareils électroménagers destinés aux loisirs, comme les téléviseurs, les chaînes stéréo, etc.

6 **Compléments d'information** : RAS

## Dictionnaire bilingue de la distribution (article type) (2)

7 **Informations linguistiques**

L'usage du terme **produits blancs** est plus fréquent en Europe qu'au Canada. L'expression **secteur du blanc** désigne l'ensemble de ces produits, mais aussi les **produits textiles de la salle de bains et de cuisine** (*household linen*).

t.c. **secteur du blanc** Voir **produits blancs**  
t.c. **secteur du blanc**(*household linen*)

8 **Contextes** :

But the European and North American markets for so-called **white goods** – appliances such as vacuum cleaners, sewing machines and dishwashers – has consolidated too much for the company to expand much further in that direction [...]. (*Wall Street Journal* #)

En revanche, les ventes de **produits blancs** (les gros et petits appareils ménagers) avancent des résultats moroses avec -0,5 % au mois de janvier, - 0,5 % au mois de février et -1,7 % en mars.

([http://www.melun.cci.fr/point\\_conjoncture/activitehtm](http://www.melun.cci.fr/point_conjoncture/activitehtm)) (#)

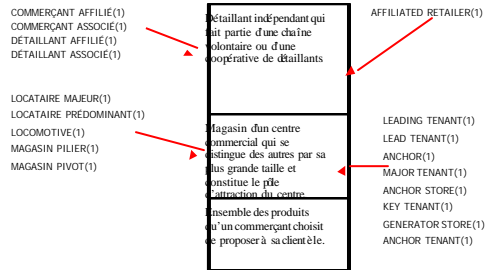
9 **Exemples**

Machines à laver, cuisinières, réfrigérateurs, etc.

## Conversion du dictionnaire en format électronique

1. Conversion dans un modèle « base de données relationnelles »
2. Systématisation des explications données sur les relations sémantiques entre termes
3. Utilisation des fonctions lexicales pour représenter les relations sémantiques entre termes.

## Modèle relationnel



## Relations sémantiques entre termes dans le dictionnaire

La **vente directe** est une des formes de **vente hors magasin** ... (énoncé apparaissant dans l'article *vente directe*).

Le **centre de liquidation** et le **magasin de faillite** sont des **magasins minimarge** (énoncé apparaissant dans l'article *magasin minimarge*).

La **vente personnelle** s'oppose à la **vente impersonnelle** (énoncé apparaissant dans l'article *vente personnelle*).

*arborer une enseigne, porter une enseigne, développer une enseigne, développement de l'enseigne, expansion de l'enseigne, prendre l'enseigne, implanter une enseigne* (liste apparaissant dans l'article *enseigne*)

L'action de **majorer** s'appelle aussi **majoration** ou **surmarquage** (énoncé apparaissant dans l'article *majoration*).

## Fonctions lexicales (Mel' cuk et al. 1988; 1995)

Formellement, une fonction lexicale (= FL) s'exprime  $f(x) = y$   
Où :  $f$  est la fonction  
 $x$  est l'argument  
 $y$  est la valeur de la fonction appliquée à l'argument  $x$

$S_0$ (prospector) = prospection  
 $S_1$ (enchère) = enchérisseur  
 $A_0$ (succursale) = succursaliste  
 $Oper_1$ (enseigne) = porter, arborer [ART ~]

## Pourquoi les FL ?

- Permettent la représentation d'une variété de relations sémantiques;
- Environ 60 FL standard
- Relations sémantiques différentes prises en compte par le même formalisme :  
synonymie, antonymie, hyperonymie, relations actanciellles, circonstanciellles, collocationnelles, etc.
- Méthode de description systématique des relations sémantiques

## Exemple (1)

VENTE, AUX ENCHÈRES, ENCHÈRE, VENTE, PAR ADJUDICATION, ENCAN<sub>m</sub>, VENTE, À LA CRIÉE (Fr.)

**Précisions sémantiques** : Le **commissaire-priseur** procède aux **enchères** qui se font à vive voix dans une **salle de vente**. Il présente l'article et demande une première **offre** ou annonce lui-même un prix initial minimal. Les acheteurs potentiels font des offres pour **enchérir**, chaque offre étant supérieure à la précédente. Le commissaire-priseur doit **adjudger** l'article au dernier **enchérisseur**. Si celui-ci n'est pas capable de payer la somme offerte, s'il a fait une **folle enchère**, l'article sera revendu. Si le prix obtenu la deuxième fois est inférieur à celui de la folle enchère, le **fol enchérisseur** devra payer la différence.

## Exemple (2)

### Enchère1

Syn(enchère1) = offre

S<sub>1</sub>(enchère1) = enchérisseur

V<sub>0</sub>(enchère1) = enchérir

Magn + AntiVer(enchère1) = folle enchère

Magn + AntiVerS<sub>1</sub>((enchère1)) = fol enchérisseur

### Enchère2

Syn(enchère2) = vente aux enchères

S<sub>1</sub>(enchère2) = commissaire-priseur

Oper<sub>1</sub>(enchère2) = procéder aux -s

S<sub>loc</sub>(enchère2) = salle de vente

## Quelques écarts et difficultés d'application des FL au dictionnaire (1)

- Interprétation lexicographique (voire terminographique) : rendre compte de relations sémantiques entre termes
- La plupart des termes décrits dans le dictionnaire monosémiques
- Problème posé par les termes complexes à sens compositionnel

*bien durable versus bien*

*ou durable versus bien*

- Les relations étaient établies à partir de l'information fournie dans le dictionnaire (travail semblable pour un dictionnaire bilingue : Fontenelle 1997)

## Quelques écarts et difficultés d'application des FL au dictionnaire (2)

- Relations non exhaustives : termes qui apparaissent dans le dictionnaire et non à faire le relevé des relations possibles entre un terme et d'autres termes du domaine

*marchandisage : mais pas marchandiser*

- Définitions analytiques (incluant et caractéristiques spécifiques) et non sous forme propositionnelle avec explicitation des actants

## Quelques écarts et difficultés d'application des FL au dictionnaire (3)

- Éternel débat : rendre compte de relations lexicales ou conceptuelles ?
- La plupart des relations dégagées dans le dictionnaire étaient de nature taxinomique ou méronymique.
- Nombreuses relations du dictionnaire ne pouvaient être représentées au moyen de FL (du moins de FL standard)
  - but visé (*locomotive : pouvoir d'attraction*)

## Plan

- **Expérience 1** : conversion d'un dictionnaire papier existant (*Dictionnaire bilingue de la distribution*, Dancette et Réthoré 2000)
  - Description du dictionnaire papier
  - Modèle relationnel
  - Modélisation des relations sémantiques au moyen des fonctions lexicales (Melcuk et al. 1988; 1995).
  - Bilan
- **Expérience 2** : **Élaboration d'un dictionnaire d'informatique**
  - **Objectifs du dictionnaire**
  - **Sélection des entrées**
  - **Distinctions sémantiques**
  - **Relations sémantiques entre termes**

## Objectifs du dictionnaire

- **Termes fondamentaux** :
  - « terme » : sens lié au domaine de l'informatique ;
  - « fondamental » : se retrouve dans de nombreux textes d'informatique (pas spécifique à une spécialisation de l'informatique) ;
- Le choix des entrées doit refléter le réseau lexical du domaine de l'informatique (séries dérivationnelles ; combinatoire, autres liens paradigmatiques).

## Contenu envisagé

- Distinguer les acceptions de formes linguistiques ayant plusieurs sens spécialisés ;
- Modélisation des liens paradigmatiques et syntagmatiques des termes faisant l'objet d'une entrée.

## Exemple

- **Dictionnaire** (amorcer1(1))

## Sélection des termes : le corpus

- A partir d'un corpus d'environ 600 000 mots

Subdivisions du corpus d'informatique	Taille des corpus	
	Nombre de textes	Nombre de mots
Initiation à la micro-informatique	8	116 821
Internet	12	102 972
Logiciel	4	70 412
Matériel	5	41 016
Programmation et réseaux	11	38 909
Systèmes d'exploitation	13	221 104
<b>Total</b>	<b>53</b>	<b>600 024</b>

## Constitution d'une liste préliminaire au moyen d'un calcul des spécificités

- Recours à TermoStat (Drouin 2003) ;
- Calcul des spécificités appliquées à l'ensemble des formes (étiquetées et lemmatisées du corpus) ;
- Comparaison de corpus ;
- Répartition des formes dans trois groupes (loi normale) :
  1. les spécificités positives (SP+) : celles dont la fréquence est plus élevée que celle observée dans le corpus de référence (nous tenons pour acquis que les termes se situent dans cette catégorie) ;
  2. les formes banales (SP0) : celles dont la fréquence est la même que celle observée dans le corpus de référence ;
  3. les spécificités négatives (SP-) : celles dont la fréquence est moins élevée que celle observée dans le corpus de référence.

## Deux méthodes

- Expérimentation et 1re évaluation (Chantal Lemay (2003)).
- Méthode 1 : comparaison du corpus d'informatique à un corpus journalistique (*Le Monde* 2001 qui comprend environ 30 millions de mots).
- Méthode 2 : subdivision du corpus d'informatique en six sous-corpus représentatifs (matériel, Internet, etc.) et à les comparer à l'ensemble des textes d'informatique.
- Fusion des listes de spécificités produites par les deux méthodes.

## Premiers résultats : Méthode 1

Forme	P. du D.	Fréquence	Valeur-test
fishior	SDC	2000	260.025
commande	SDC	1002	204.740
option	SDC	1400	192.328
utilisateur	SDC	1117	188.477
configuration	SDC	845	167.307
utiliser	VB	1936	162.02
repertoire	SBC	1003	161.708
système	SBC	2699	153.668
disquette	SBC	609	152.979
ordinateur	SBC	1283	148.431
touche	SBC	865	140.484
logiciel	SDC	1100	138.648
imprimante	SDC	537	137.022
clacq	SDC	1093	137.022
répertoire	SDC	1240	129.009
windows	SDP	613	125.77
cavier	SDG	579	125.066
caractère	SBC	1096	122.24
recommander	ADJPAP	516	118.081
lieux	SBP	442	117.004
votre	DT	1365	116.261
bit	SBC	382	110.826
paramètre	SDC	465	110.72
interface	SBC	412	100.066

## Premiers résultats : Méthode 2

Forme	P. du D.	Fréquence	Valeur-test	Sous-corpus
internet	CDC	792	11,4712	CG2
mémoire	CDC	737	39,4842	CG1
option	CDC	1253	29,3491	CG6
substantif	CDC	342	29,7559	CG3
impressionne	CDC	269	25,9919	CG4
papier	SBC	104	25,0858	SC4
fishier	CDC	3299	23,9704	CG6
implémentation	SBC	35	22,65	SC5
rapport	SBP	44	22,101	SC4
recommander	ADJPAP	496	21,4903	SC6
algorithme	SBC	61	20,2627	SC5
synchronisation	CDC	63	20,8597	CG6
renseignement	SBC	36	19,994	SC5
noyau	CDC	430	19,9906	CG6
contrat	CDC	44	19,9522	CG5
laser	CDC	62	19,973	CG4
électronique	ADJ	292	19,9512	CG2
linux	SBP	422	19,6613	SC6
france	SBP	154	19,6566	SC2
étendu	CDC	75	19,1454	CG3
configuration	CDC	660	19,8863	CG6
implément	ADJ	52	18,9279	SC4
feuille	CDC	67	19,8894	CG4
réseau	CDC	292	19,7129	CG6

## Sélection des termes : critères lexico-sémantiques (1)

- Les unités extraites désignent une entité (matériel, logiciel, entité de représentation, unité de mesure ou un animé) du domaine de l'informatique (ex. *archive, carte, compilateur, programme*) ;
- S'il s'agit d'unités prédictives – verbes, nominalisations, adjectifs, etc. –, elles sont extraites si les actants renvoient à des entités du critère a. (ex. *charger : l'utilisateur charge un logiciel en mémoire*). Toutefois, la même unité prédictive peut se combiner avec des actants non spécialisés; si elle revêt le même sens avec ces autres actants, elle est éliminée.

## Sélection des termes : critères lexico-sémantiques (2)

- Si l'il s'agit de dérivés morphologiques, ils sont sélectionnés s'ils sont sémantiquement apparentés à un terme sélectionné en fonction des critères a. ou b. (ex. *amorcer: amorçable, amorcer, réamorcer, etc.*; *archiver : archivage, archive, etc.*).
- Si l'il s'agit d'une unité entrant dans une relation paradigmatique (autre qu'une relation morphologique déjà identifiée en c.) avec un terme sélectionné en fonction des critères a., b. ou c. (ex. *coller, couper, copier; hyperlien, lien*), elle est extraite.

\*\* Fréquence et répartition.

## Évaluation 2 (lettres A, C et P)

Nombre total de SP+ générées automatiquement	Nombre de SP+ sélectionnées	Précision
1047	553	52,82 %
Nombre de mots lexicaux identifiés automatiquement moins les noms propres et les participes passés)	Nombre de SP+ sélectionnées	Précision
852	472	55,47 %

## Exemples d'unités acceptées et rejetées

Unité	P. de d.	Clinguoy	Collin	Sélectionné	Méthode
abome	SBC	oui	oui	oui	1 et 2
application	SBC	oui	oui	oui	1 et 2
abonnement	VB	non	non	oui	1 et 2
abonner	VB	non	non	oui	1
abonné	SBC	oui	oui	oui	1 et 2
accessibilité	SBC	oui	non	oui	1
accessible	ADJ	oui	oui	oui	1
algorithme	SBC	oui	oui	oui	1 et 2
algorithmique	ADJ	non	non	oui	1 et 2
automatique	ADJ	oui	oui	oui	1
automatique	ADV	non	non	oui	1
automatisé	SBC	oui	oui	oui	1
exercice	SBC	oui	oui	oui	1 et 2
chemin	SBC	oui	oui	oui	2
chavir	SBC	oui	oui	oui	1 et 2
partage	SBC	oui	oui	oui	1 et 2
partageable	ADJ	oui	non	oui	1
partage	ADJPAP	non	non	oui	1 et 2
affir	VB	oui	oui	non	2
absolument	ADV	non	non	non	2
accord	SBC	non	oui	non	1
affirmative	SBC	non	non	non	1
agir	VB	non	non	non	1
ainsi	ADV	non	non	non	1
alcool	SBC	non	oui	non	2
aune	ADJ	non	non	non	1
avantage	SBC	non	non	non	1
capable	ADJ	non	oui	non	1 et 2

## Distinctions sémantiques : critères lexico-sémantiques (1)

- Cooccurrence compatible (Mel'cuk et al. 1995 : 64-65): Si les cooccurents peuvent être combinés et donner lieu à une phrase acceptable, un seul sens est identifié.  
*exécuter Word, une application ou le système d'exploitation*  
*initialiser le disque dur ou la disquette*
- Cooccurrence différentielle (Mel'cuk et al. 1995: 64-65) : Si les cooccurents sont combinés et donnent lieu à une phrase inacceptable, plusieurs sens sont dégagés.  
*adresser la mémoire adresser un message*  
*\*adresser la mémoire et le message*  
*Formater un document formater un disque*  
*\*formater un document et un disque*  
*\*initialiser l'ordinateur et la disquette*

Cruse (1986), L'Homme (1998, 2003), Mel'cuk et al. (1995)

## Distinctions sémantiques : critères lexico-sémantiques (2)

- c) Remplacement par un synonyme : Lorsqu'un synonyme peut remplacer toutes les occurrences d'une unité lexicale et donner lieu au même sens, un seul sens est identifié; si le synonyme remplace une partie des occurrences mais pas toutes, plusieurs sens sont dégagés.

*exécuter un logiciel*                      *exécuter une tâche*  
*\*accomplir un logiciel*                      *accomplir une tâche*

- d) Dérivation morphologique différentielle : Lorsque deux séries de dérivés morphologiques peuvent être dégagées pour une même unité lexicale, deux sens sont identifiés.

*programmer une application*                      *programmer une mémoire*  
*? application programmable*                      *mémoire programmable*  
*installer un logiciel*                      *installer un ordinateur*  
*désinstaller un logiciel*                      *?désinstaller un ordinateur*

## Distinctions sémantiques : critères lexico-sémantiques (3)

- e) Autres liens paradigmatiques : Lorsque deux séries d'unités lexicales reliées sur le plan paradigmatique peuvent être dégagées pour une même unité lexicale, deux sens sont identifiés.

*page* : *page Web, lien, adresse, portail*  
*page* : *document, format*

*copier (quelque chose dans la mémoire tampon)* : *couper, coller*  
*copier (un fichier sur disque)* : *supprimer, effacer*

## Description de la structure actancielle

1. Nombre d'actants : actants nécessaires à la description du sens  
*installer(1)* : X installe Y  
*installer(2)* : X installe Y sur Z
2. Rôle sémantique des actants (agent, instrument, destination, etc.)  
*installer(1)* : agent installe patient  
*installer(2)* : agent installe patient sur destination
3. Regroupement des actants en classes sémantiques  
*installer2* : agent(utilisateur) installe patient(application) sur destination(support permanent)  
*connecter1* : agent(utilisateur) se connecte à destination(réseau)  
*tourner(1)* : agent(application) tourne sur destination(ordinateur)

## Structures actancielles régulières

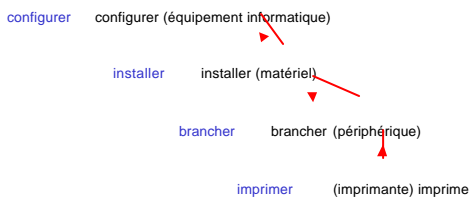
**programmer(1)**  
 agent(utilisateur) programme patient(mémoire)

**programmation(1a)**  
 programmation de patient(mémoire) par agent(utilisateur)

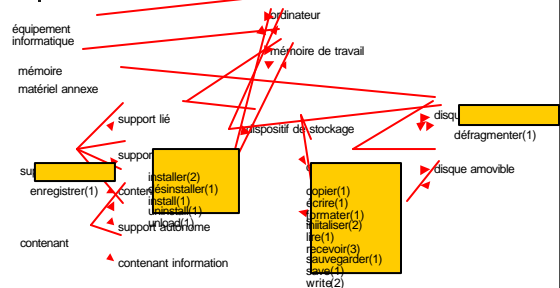
**programmable(1)**  
 patient(mémoire) est programmable

**reprogrammer(1)**  
 agent(utilisateur) reprogramme patient(mémoire)

## Classes sémantiques en informatique



## Hierarchisation des classes sémantiques



## Description des relations paradigmatiques et syntagmatiques

Fonctions lexicales et version vulgarisée (DiCo)

**Ordinateur(1)** : ordinateur [utilisé par] agent(utilisateur) [pour intervenir sur] patient(données)

- Terme plus général (Syn) = *machine, appareil*
- Quasi-synonyme (Qsyn) = *calculateur*
- Terme plus spécifique (Spec) = *micro-ordinateur, portable*
- Terme plus spécifique (Spec) = *client, serveur*
- Une collection (Mult) = *réseau d' -s*
- Fonctionne (Fact0) = *~ tourne*
- L'agent prépare (Prepar) = *configurer l' -*
- Quelqu'un commence à faire fonctionner l'ordinateur (CausFact0) = *démarrer l' -, initialiser l' -*

## Current state of the dictionary

- Approx. 1 200 entrées (termes définis jusqu'à maintenant comme des entrées valables);
- Approx. 450 structure actancielle décrites et validées;
- Approx. 1 500 fonctions lexicales.

## Exemples

### Dictionnaire

- Virus
- Logiciel1
- Ordinateur
- Exécuter1, exécuter2, exécuter3

## La suite ....

- Stabiliser la macrostructure
- Mise au point d'un modèle de définition intégrant les actants sémantiques (à la LEC)
- Mise au point de stratégies de descriptions pour des termes appartenant à une même classe sémantique;
- Éventuellement, s'aligner plus directement aux descriptions du DiCo.

## Pistes dégagées par ces deux projets

D'une manière générale, des modèles de sémantique lexicale formelle et la lexicologie explicative et combinatoire, en particulier :

- sont utiles pour décrire les termes;
- proposent au terminologue un cadre pour appréhender les données en corpus spécialisé.;
- entraînent une modulation sensible de certains principes terminologiques (notamment, l'accent mis sur le concept et les relations conceptuelles).