Une approche linguistique de l'évaluation des ressources extraites par analyse distributionnelle automatique

Thèse soutenue le mercredi 10 juillet 2013
par François Morlane-Hondère

CLLE-ERSS, Université de Toulouse

Composition du jury :
Pierre-André Buvet - Maître de conférences HDR, Université Paris 13/LDI (examinateur)
Béatrice Daille - Professeur, Université de Nantes/LINA (rapporteur)
Cécile Fabre - Professeur, Université Toulouse II/CLLE-ERSS (directrice)
Nabil Hathout - DR, CNRS et Université Toulouse II/CLLE-ERSS (examinateur)
Alain Polguère - Professeur, Université de Lorraine/ATILF (rapporteur)

Plan

1. L'analyse distributionnelle automatique
2. Évaluation globale : recouvrement bases distributionnelles/lexiques
3. Évaluation locale : une approche linguistique
   - Effets du filtrage du DES à l'aide des voisins
   - Une description syntagmatique et paradigmatique de l'antonymie
   - Étude de la substituabilité des hyperonymes dans les voisins
   - Étude des manifestations de la méronymie dans les voisins
4. Apports et perspectives

Plan

1 L'analyse distributionnelle automatique

2 Évaluation globale : recouvrement bases distributionnelles/lexiques

3 Évaluation locale : une approche linguistique
- Effets du filtrage du DES à l'aide des voisins
- Une description syntagmatique et paradigmatique de l'antonymie
- Étude de la substituabilité des hyperonymes dans les voisins
- Étude des manifestations de la méronymie dans les voisins

4 Apports et perspectives

L'hypothèse distributionnelle

- Distributionnalisme (Bloomfield 1935)
- Héritage de Zellig. S. Harris

Harris (1970), p. 786

[...] if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. **In other words, difference of meaning correlates with difference of distribution.**

- Mise au jour de relations de sens

Automatisation du principe distributionnel

- Enjeu : constitution automatique de thesaurus
- $\sim$ milieu des années 90 (Grefenstette 1994)
  - disponibilité de bases textuelles au format électronique
  - développement d'analyseurs syntaxiques opérationnels

$\longrightarrow$ mise en œuvre du principe distributionnel à grande échelle

- Diversification des types de textes abordés :
  - des textes de spécialité...
  - ...à la "langue générale" (textes journalistiques, littéraires, encyclopédiques, Web...)

Capter les relations lexicales

- Exemples issus des Voisins de Wikipédia :

couple	exemples de contextes
livre/ouvrage	*préfacer*_OBJ, *dédicacer*_OBJ...
premier/dernier	*affronter*_SUJ, *s'intituler*_SUJ...
bière/boisson	*boire*_OBJ, *consommer*_OBJ...
avoine/maïs	*farine*_DE, *cultiver*_OBJ...
moteur/voiture	*constructeur*_DE, *panne*_DE...
piano/art martial	*apprendre*_OBJ, *maîtriser*_OBJ...
cerveau/édifice	*structure*_DE, *taille*_DE...
arbre/usine	*produire*_SUJ, *fournir*_SUJ...

Capter les relations lexicales

- Exemples issus des Voisins de Wikipédia :

couple	exemples de contextes	
livre/ouvrage	*préfacer*_OBJ, *dédicacer*_OBJ...	→ **synonymie**
premier/dernier	*affronter*_SUJ, *s'intituler*_SUJ...	→ **antonymie**
bière/boisson	*boire*_OBJ, *consommer*_OBJ...	→ **hyperonymie**
avoine/maïs	*farine*_DE, *cultiver*_OBJ...	→ **co-hyponymie**
moteur/voiture	*constructeur*_DE, *panne*_DE...	→ **méronymie**
piano/art martial	*apprendre*_OBJ, *maîtriser*_OBJ...	→ **?**
cerveau/édifice	*structure*_DE, *taille*_DE...	→ **?**
arbre/usine	*produire*_SUJ, *fournir*_SUJ...	→ **?**

- Relations non typées

- Nature des relations extraites ? Proportions ? Modalités d'extraction ?

→ caractérisation des ressources distributionnelles ?

→ frein à l'exploitation de l'analyse distributionnelle automatique (ADA)

Une approche linguistique de l'évaluation

- Approches *globales* : comparaison à des lexiques, évaluation par la tâche, test du TOEFL...
→ portée des résultats limitée

vs

- Approche *locale* : retour aux contextes d'apparition
→ problème de la masse

- Combinaison des deux approches :
  1. comparaison des bases distributionnelles à des lexiques
  2. analyse des propriétés distributionnelles de sous-ensembles de couples qui apparaissent – ou pas – dans l'intersection

Plan

1 L'analyse distributionnelle automatique

2 Évaluation globale : recouvrement bases distributionnelles/lexiques

3 Évaluation locale : une approche linguistique
- Effets du filtrage du DES à l'aide des voisins
- Une description syntagmatique et paradigmatique de l'antonymie
- Étude de la substituabilité des hyperonymes dans les voisins
- Étude des manifestations de la méronymie dans les voisins

4 Apports et perspectives

Ressources (1) : bases distributionnelles

- Ressources développées au CLLE-ERSS

- Corpus : encyclopédique (Wikipédia), journalistique (Le Monde) et littéraire (Frantext)

Mise en œuvre de l'analyse distributionnelle

1. analyse syntaxique du corpus par Syntex (Bourigault 2007)
2. calcul de la similarité entre les vecteurs distributionnels par Upéry (Bourigault 2002), score de Lin (1998) $\geq 0{,}1$

$\Downarrow$

	voisins de *		
corpus	Wikipédia	Le Monde	Frantext
taille (en nombre de couples)	3 922 657	5 525 480	792 356

- Consultables sur la plate-forme REDAC (http://redac.univ-tlse2.fr/)

Ressources (2) : lexiques

- Dictionnaire Électronique des Synonymes (DES) (Manguin *et al.* 2004)
  - CRISCO, Université de Caen
  - Compilation des synonymes contenus dans 7 dictionnaires
  - 389 182 paires de noms (52 %), verbes (26 %) et adjectifs (22 %)
- JeuxDeMots (Lafourcade 2007)
  - LIRMM, Université de Montpellier
  - Ressource collaborative enrichie *via* un jeu en ligne
  - Relations *classiques* (synonymie, antonymie, hyperonymie, méronymie) et *non classiques* (*chose/lieu*, *action/instrument*, etc.)

Portée de l'étude

- Approche nécessairement bornée
- Une étude des relations *classiques* (Morris et Hirst 2004)
  - synonymie
  - antonymie
  - hyperonymie
  - méronymie
- Biais des lexiques

Un recouvrement global faible

- Mesure du recouvrement entre les VDW et le DES/JDM (lexique commun)

	DES	JeuxDeMots		
		antonymie	hyperonymie	méronymie
Nb. de couples total	99 235	4838	17 020	15 912
Nb. de couples partagés	41 718	1770	6314	5380
Recouvrement (%)	42	37	37	34

- Point de départ d'une série d'approches qualitatives

Plan

1 L'analyse distributionnelle automatique

2 Évaluation globale : recouvrement bases distributionnelles/lexiques

3 Évaluation locale : une approche linguistique
- Effets du filtrage du DES à l'aide des voisins
- Une description syntagmatique et paradigmatique de l'antonymie
- Étude de la substituabilité des hyperonymes dans les voisins
- Étude des manifestations de la méronymie dans les voisins

4 Apports et perspectives

Éléments de méthodologie

- 4 protocoles distincts, degré de complexité croissant

- Relations extraites par l'ADA *vs* relations non extraites
→ analyse d'échantillons

- Filtrage par seuil de productivité (Weeds 2003)
→ seuls sont étudiés les couples composés de mots aux productivités comparables

Plan

1 L'analyse distributionnelle automatique

2 Évaluation globale : recouvrement bases distributionnelles/lexiques

3 Évaluation locale : une approche linguistique
- Effets du filtrage du DES à l'aide des voisins
- Une description syntagmatique et paradigmatique de l'antonymie
- Étude de la substituabilité des hyperonymes dans les voisins
- Étude des manifestations de la méronymie dans les voisins

4 Apports et perspectives

Problématique et méthodologie

- Recouvrement modéré VDW/DES
→ **42 %** des synonymes sont extraits dans les VDW

- Synonymes : haut degré de substituabilité

Comment s'explique le non-repérage de la plupart des synonymes par l'ADA ?

- Échantillon : 30 noms, 30 verbes et 30 adjectifs dont les synonymes ont été les moins extraits
  - noms : *éclat* (29), *remède* (21), *espérance* (20), *flamme* (20)...
  - verbes : *exciter* (29), *vider* (22), *coller* (22), *ôter* (20)...
  - adjectifs : *fou* (30), *doux* (27), *rude* (27), *ferme* (25)...

Causes du non-repérage des synonymes

- Phénomènes identifiés :
  - la polysémie (acceptions spécialisées, emplois métaphoriques)
  - la connotation (énonciative, stylistique)
  - la dénotation périphérique (absence/présence/différence de traits)

- Exemple de décalage : *fraternité*

Contextes : *rejoindre*_OBJ, *membre*_DE, *fondateur*_DE...
→ voisins : *phalange*, *parti ouvrier*, *congrégation*, *rassemblement*...

$\neq$

Synonymes : *charité*, *sympathie*, *confiance*, *solidarité*, *générosité*...

Approche inter-corpus

- Influence du type de corpus sur l'extraction des synonymes ?

- Comparaison des synonymes extraits : corpus encyclopédique (VDW) vs journalistique (VDLM)

Résultats

Nombre moyen de synonymes extraits par les voisins :

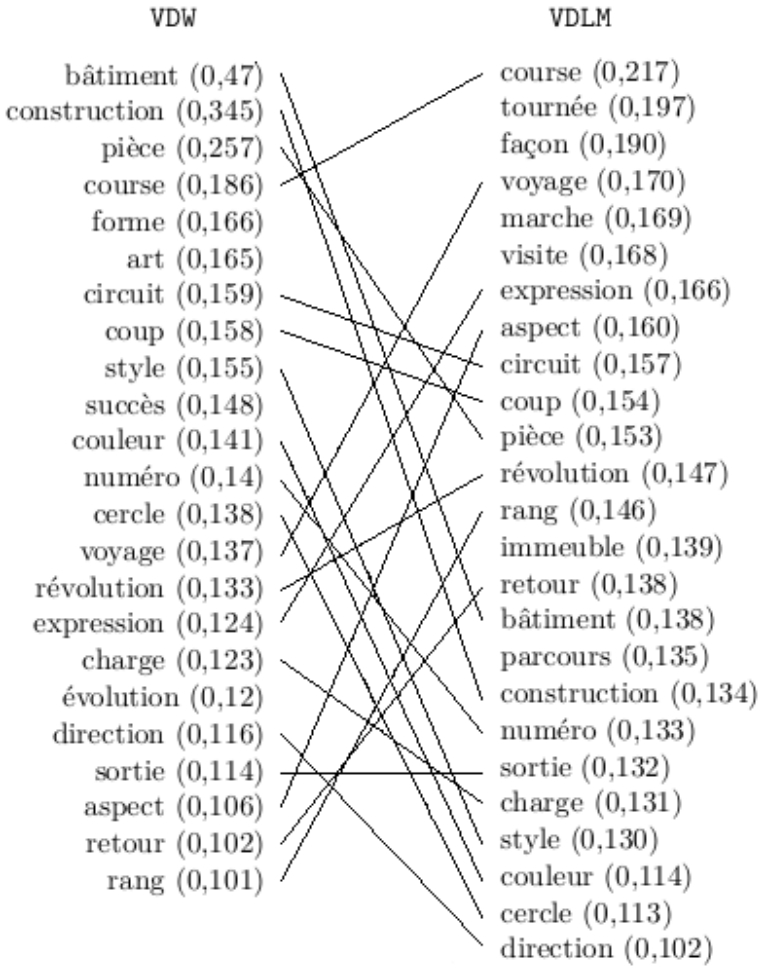
	VDW	VDLM
synonymes extraits	11,4	11,4
dont communs	10	

Les VDW et les VDLM ont tendance à extraire les mêmes synonymes.

→ **changement de point de vue**

Classement par score de similarité

- Tri des synonymes de *tour* extraits par les VDW et les VDLM par degré de similarité décroissant



Classement par score de similarité

■ Tri des synonymes de *tour* extraits par les VDW et les VDLM par degré de similarité décroissant

	VDW	VDLM
#1	*bâtiment* (0,470)	*course* (0,217)
#2	*construction* (0,345)	*tournée* (0,197)
#3	*pièce* (0,257)	*façon* (0,190)
#4	*course* (0,186)	*voyage* (0,170)
#5	*forme* (0,166)	*marche* (0,169)
#6	*art* (0,165)	*visite* (0,168)
#7	*circuit* (0,159)	*expression* (0,166)
#8	*coup* (0,158)	*aspect* (0,160)
#9	*style* (0,155)	*circuit* (0,157)
#10	*succès* (0,148)	*coup* (0,154)

Résultats

- Mise au jour de phénomènes linguistiques empêchant le repérage de certains synonymes

- Approche inter-corpus : démonstration du potentiel de l'ADA pour la réorganisation des synonymes du DES

Plan

1 L'analyse distributionnelle automatique

2 Évaluation globale : recouvrement bases distributionnelles/lexiques

3 Évaluation locale : une approche linguistique
- Effets du filtrage du DES à l'aide des voisins
- Une description syntagmatique et paradigmatique de l'antonymie
- Étude de la substituabilité des hyperonymes dans les voisins
- Étude des manifestations de la méronymie dans les voisins

4 Apports et perspectives

Problématique et méthodologie

- Redéfinition de la relation d'antonymie comme une relation syntagmatique (Fellbaum 1995, Jones 2002)

- Combiner patrons et analyse distributionnelle automatique
→ capter les fonctionnements syntagmatiques et paradigmatiques

Quels sont les apports de l'ADA à une méthode d'extraction des couples d'antonymes basée dur des patrons lexico-syntaxiques ?

Protocole

1 – Adaptation des patrons antonymiques de Jones (2002)

patron	exemple
X *ou* **Y**	Cette connexion peut être **temporaire** *ou* **définitive**
X *comme* **Y**	Elle peut **s'arrêter** *comme* **continuer**
ni **X** *ni* **Y**	Il n'est donc *ni* **explicite** *ni* **implicite**
de/depuis **X** *à/jusqu'à* **Y**	La région s'étend *du* **nord** *au* **sud** sur 350 km
plus/plutôt/moins/ autant/aussi (bien) **X** *que* **Y**	Il se déguste *aussi bien* **chaud** *que* **froid**

2 – Projection des patrons sur le corpus Wikipédia
→ extraction de 75 389 couples (dont 28 318 d'adjectifs)
→ résultats ambigus :

"Le terme, néanmoins, reflète des connotations [*plus* **humaines** *qu'***administratives**]."

Protocole

3 – Comparaison des couples extraits avec les VDW
$\longrightarrow$ filtrage des oppositions occasionnelles

4 – Évaluation

- comparaison avec un lexique en ligne (dictionnaire d'antonymes du CRISCO)
- sollicitation de locuteurs *via* des questionnaires

Résultats

- Les voisins :
  - n'apparaissent pas significativement dans le lexique
  - ne sont pas davantage identifiés comme des antonymes par les locuteurs

- Limites du protocole
  - couples antonymiques absents du lexique
    → *nuisible/bénéfique*, *hétérosexuel/homosexuel*, *acide/basique*...
  - difficulté de la tâche de jugement de l'antonymie
    → taux d'accord fluctuant (kappa entre 0,27 et 0,64)

Plan

1 L'analyse distributionnelle automatique

2 Évaluation globale : recouvrement bases distributionnelles/lexiques

3 Évaluation locale : une approche linguistique
- Effets du filtrage du DES à l'aide des voisins
- Une description syntagmatique et paradigmatique de l'antonymie
- Étude de la substituabilité des hyperonymes dans les voisins
- Étude des manifestations de la méronymie dans les voisins

4 Apports et perspectives

Problématique et méthodologie

- Faible recouvrement avec JeuxDeMots
- Substituabilité des hypo/hyperonymes ?

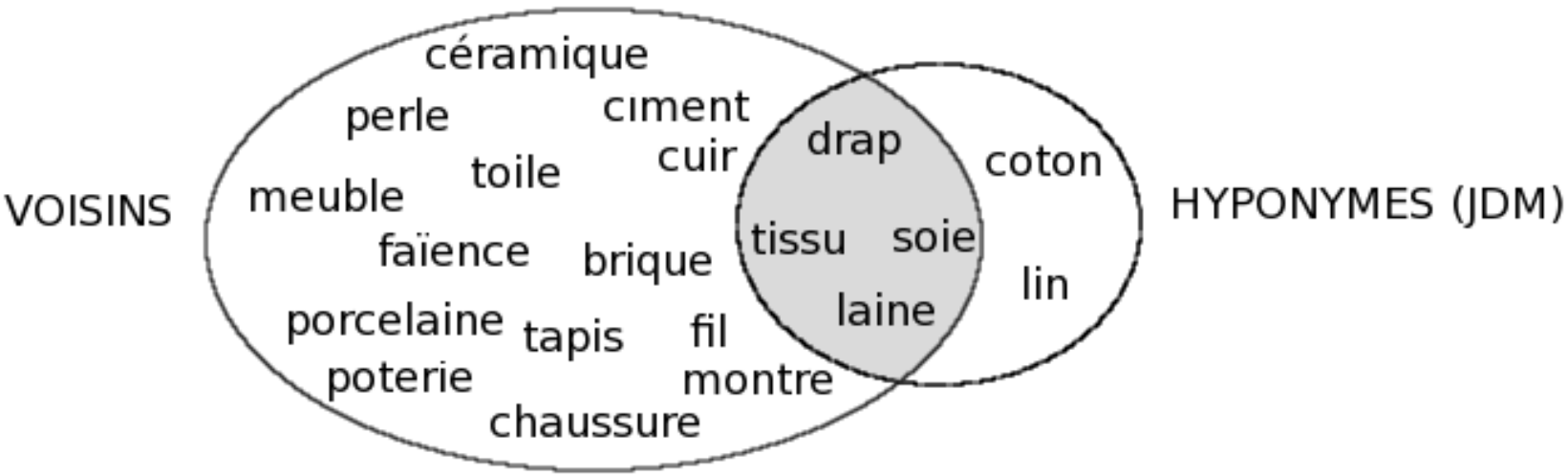
Est-ce que la catégorisation des hyponymes dans les textes correspond à une catégorisation *en langue* ?

- Échantillonage basé sur le nombre d'hyponymes extraits par hyperonyme
→ calcul du rappel

$$rappel = \frac{nombre\ d'hyponymes\ captés}{nombre\ total\ d'hyponymes}$$

Calcul du rappel des hyperonymes

- Exemple : recouvrement des voisins et des hyponymes du mot *étoffe*



- $rappel_{étoffe} = \frac{4}{6} = 0,67$

✓ : voisin
✗ : présent dans le lexique, non voisin
∅ : mot absent du lexique des voisins ou filtré (non pris en compte dans le rappel)

émotion	VDW	VDLM	VDF
bienfait		✗	✗
bonheur	✓	✓	✓
colère	✓	✓	✓
enthousiasme	✓	✓	✗
envie	✓	✓	✓
joie	✓	✓	✓
passion	✓	✓	✓
peur	✓	✓	✓
plaisir	✓	✓	✓
rire		✗	✗
sensation	✓	✓	✓
souvenir	✗	✓	✓
tendresse		✓	✓
tristesse	✓	✗	✓
rappel	0,91	0,79	0,79
rappel moyen	0,83		
écart type	0,06		

engin	VDW	VDLM	VDF
bus	✓	✗	✗
camion	✓	✓	
locomotive	✓		✗
métro	✗	✗	
ordinateur	✓	✓	
outil		✓	
pelle			✗
planeur	✓		
satellite	✓	✓	
train		✓	
voiture		✓	
wagon	✓	✗	
rappel	0,87	0,33	0
rappel moyen	0,40		
écart type	0,36		

Une analyse en deux temps

- Les hyperonymes se définissent selon deux critères :
  - le rappel moyen
    $\longrightarrow$ proportion d'hyponymes repérés par les voisins
  - l'écart type
    $\longrightarrow$ influence du genre du corpus sur le repérage des hyponymes

- Deux échantillons d'hyperonymes

Hyperonymes dont les hyponymes sont bien repérés dans les trois bases

- Catégorisation opérée dans les trois corpus conforme à celle du lexique

hyperonyme	rappel	écart type	hyperonyme	rappel	écart type
parent	1	0	*femme*	0,87	0,02
peur	1	0	*émotion*	0,85	0,05
enfant	0,95	0,07	*salle*	0,83	0,13
soldat	0,93	0,1	*sentiment*	0,8	0,06
type	0,93	0,09	*prince*	0,77	0,21
homme	0,93	0,05	*habit*	0,77	0,15
petit	0,92	0,12	*son*	0,77	0,07
livre	0,92	0,12	*métal*	0,75	0,18
officier	0,92	0,12	*pensée*	0,74	0,07
artiste	0,91	0,07	*corps*	0,74	0,05
chef	0,89	0,16	*école*	0,72	0,16
employé	0,88	0,09	*temps*	0,72	0,06
trace	0,87	0,19	*vue*	0,69	0,2
ouvrage	0,87	0,19	*visage*	0,68	0,14
aventure	0,87	0,19	*personne*	0,68	0,05

Hyperonymes dont les hyponymes sont bien repérés dans les trois bases

- Catégorisation opérée dans les trois corpus conforme à celle du lexique

hyperonyme	rappel	écart type	hyperonyme	rappel	écart type
parent	1	0	*femme*	0,87	0,02
peur	1	0	*émotion*	0,85	0,05
enfant	0,95	0,07	*salle*	0,83	0,13
soldat	0,93	0,1	*sentiment*	0,8	0,06
type	0,93	0,09	*prince*	0,77	0,21
homme	0,93	0,05	*habit*	0,77	0,15
petit	0,92	0,12	*son*	0,77	0,07
livre	0,92	0,12	*métal*	0,75	0,18
officier	0,92	0,12	*pensée*	0,74	0,07
artiste	0,91	0,07	*corps*	0,74	0,05
chef	0,89	0,16	*école*	0,72	0,16
employé	0,88	0,09	*temps*	0,72	0,06
trace	0,87	0,19	*vue*	0,69	0,2
ouvrage	0,87	0,19	*visage*	0,68	0,14
aventure	0,87	0,19	*personne*	0,68	0,05

Hyperonymes dont les hyponymes sont mal repérés dans les trois bases

- Catégorisation opérée dans les corpus non conforme à celle du lexique
- Causes identifiées : polysémie, facettes, emplois en usage *vs* mention, relations non hyperonymiques
- Exemple de recatégorisation : *plante*

plante	*oignon*
*liste*_DE	*ajouter*_OBJ
*assigner*_OBJ	*couper*_OBJ
*famille*_DE	*hacher*_OBJ
*jardin*_DE	*éplucher*_OBJ
*espèce*_DE	*pelure*_DE
*genre*_DE	*soupe*_À
*cultiver*_COMME	*culture*_DE
*cultiver*_OBJ	*variété*_DE
*partie*_DE	*cuire*_OBJ
*utiliser*_OBJ	*tarte*_À

Résultats

- Calcul du rappel et de l'écart type des hyperonymes
  → propotion d'hyponymes repérés, stabilité du repérage

- Catégorisation du corpus *vs* catégorisation du lexique :
  - conformité
    → cas des noms d'humains
  - décalage
    → identification de quatre phénomènes qui entrainent la recatégorisation

Plan

1 L'analyse distributionnelle automatique

2 Évaluation globale : recouvrement bases distributionnelles/lexiques

3 Évaluation locale : une approche linguistique
- Effets du filtrage du DES à l'aide des voisins
- Une description syntagmatique et paradigmatique de l'antonymie
- Étude de la substituabilité des hyperonymes dans les voisins
- Étude des manifestations de la méronymie dans les voisins

4 Apports et perspectives

Problématique et méthodologie

- Extraction automatique : application du principe distributionnel ?
  - *voiture*/*automobile* (synonymie)
  - *voiture*/*véhicule* (hyponymie)
  - *moteur*/*voiture* (méronymie)

→ repérées dans des proportions comparables

Quels types de méronymes sont extraits par l'analyse distributionnelle ?

- Double annotation des méronymes de JeuxDeMots
  1. annotation des sous-relations (typologie de Winston *et al.* (1987))
     → résultats inexploitables
  2. **annotation en classes sémantiques**

Annotation sémantique des méronymes

- Hypothèses :
  1. il n'y a qu'une forme de méronymie, seule change la nature sémantique des méronymes (Murphy 2003)
  2. la nature sémantique des méronymes influe sur leurs chances d'être extraits par l'ADA

- Annotation sémantique semi-automatique

maison/*village*	$\longrightarrow$ BÂTIMENT/LIEU
mouton/*troupeau*	$\longrightarrow$ ANIMAL/COLLECTION

$\longrightarrow$ exploitation de l'arborescence du Princeton WordNet (par traduction)

Disparités dans le repérage des couples de classes

- Couples homogènes (64,1 %) *vs* hétérogènes (50,6%)

classe méro.	classe holo.	nb. de couples	% voisins
SOUL	SOCIAL GROUP	54	96,2
LOCATION	LOCATION	19	90,0
EVENT	EVENT	29	85,3
STRUCTURE	LOCATION	94	82,6
STRUCTURE	STRUCTURE	12	82,5
SOUL	STRUCTURE	25	66,7
INSTRUMENTATION	STRUCTURE	11	51,9
SUBSTANCE	SUBSTANCE	27	51,2
INSTRUMENTATION	INSTRUMENTATION	82	41,5
LIVING PART	PLANT LIFE	76	41,2
LIVING PART	LIVING PART	62	33,9
SUBSTANCE	INSTRUMENTATION	17	33,3
LIVING PART	SOUL	54	18,4
LIVING PART	FAUNA	41	14,6

Disparités dans le repérage des couples de classes

- Couples homogènes (64,1 %) *vs* hétérogènes (50,6%)

classe méro.	classe holo.	nb. de couples	% voisins
SOUL	SOCIAL GROUP	54	96,2
LOCATION	LOCATION	19	90,0
EVENT	EVENT	29	85,3
STRUCTURE	LOCATION	94	82,6
STRUCTURE	STRUCTURE	12	82,5
SOUL	STRUCTURE	25	66,7
INSTRUMENTATION	STRUCTURE	11	51,9
SUBSTANCE	SUBSTANCE	27	51,2
INSTRUMENTATION	INSTRUMENTATION	82	41,5
LIVING PART	PLANT LIFE	76	41,2
LIVING PART	LIVING PART	62	33,9
SUBSTANCE	INSTRUMENTATION	17	33,3
LIVING PART	SOUL	54	18,4
LIVING PART	FAUNA	41	14,6

Exemples de couples homogènes

Couple de classes les mieux captés : LIEU/LIEU

Allemagne/Europe, *place/village*, *commune/canton*

- sujet de verbes de localisation spatiale : *se situer*, *border*, *s'étendre*
- complément de N : *limite de* ∼, *abords de* ∼, *environs de* ∼
- prépositions caractéristiques : *au sud de*, *au centre de*

Couples moins bien captés : PARTIE DU CORPS/PARTIE DU CORPS

chair/doigt, *muscle/bras*, *peau/visage*
→ propriétés distributionnelles moins stables :

- *chair* : *manger*_OBJ, *consommer*_OBJ, *morceau*_DE
- *doigt* : *pointer*_OBJ, *muscle*_DE, *lever*_OBJ

→ classe moins bien circonscrite (organes, membres, etc.)

Disparités dans le repérage des couples de classes

- Couples homogènes (64,1 %) *vs* hétérogènes (50,6%)

classe méro.	classe holo.	nb. de couples	% voisins
SOUL	SOCIAL GROUP	54	96,2
LOCATION	LOCATION	19	90,0
EVENT	EVENT	29	85,3
STRUCTURE	LOCATION	94	82,6
STRUCTURE	STRUCTURE	12	82,5
SOUL	STRUCTURE	25	66,7
INSTRUMENTATION	STRUCTURE	11	51,9
SUBSTANCE	SUBSTANCE	27	51,2
INSTRUMENTATION	INSTRUMENTATION	82	41,5
LIVING PART	PLANT LIFE	76	41,2
LIVING PART	LIVING PART	62	33,9
SUBSTANCE	INSTRUMENTATION	17	33,3
LIVING PART	SOUL	54	18,4
LIVING PART	FAUNA	41	14,6

Exemples de couples hétérogènes

Couple de classes les mieux captés : HUMAIN/GROUPE SOCIAL

- *directeur/entreprise* : *détenir*_SUJ, *gérer*_SUJ
- *fils/famille* : *s'exiler*_SUJ, *bâtir*_SUJ
- *musicien/orchestre* : *accompagner*_SUJ, *enregistrer*_SUJ

→ phénomène de métonymie

Couples de classes moins bien captés : PARTIE DU CORPS/ANIMAL

écaille/lézard, *patte/chat*, *queue/lion*...

- *écaille* : *recouvrir*_DE, *porter*_OBJ, *armure*_DE
- *lézard* : *morsure*_DE, *espèce*_DE, *vivre*_SUJ

→ partie inanimée *vs* être animé

Résultats

- Annotation sémantique des couples de méronymes de JDM
- Influence de la nature des méronymes sur leur repérage
- Relation homogénéité/repérage par l'ADA non systématique :

→ repérage des couples homogènes inégal

→ différence de classes sémantiques mais proximité distributionnelle

- Influence de phénomènes linguistiques sur la distribution des méronymes

→ fonctionnements métonymiques

Plan

1 L'analyse distributionnelle automatique

2 Évaluation globale : recouvrement bases distributionnelles/lexiques

3 Évaluation locale : une approche linguistique
- Effets du filtrage du DES à l'aide des voisins
- Une description syntagmatique et paradigmatique de l'antonymie
- Étude de la substituabilité des hyperonymes dans les voisins
- Étude des manifestations de la méronymie dans les voisins

4 Apports et perspectives

Apports

- Originalité de l'approche linguistique
- Principaux résultats :
  - décalage entre les relations extraites et les relations *en langue*
  - influence de phénomènes linguistiques sur la distribution des mots
  - nécessité de connaître la nature du corpus pour la caractérisation des résultats fournis par l'ADA
- Destinataires :
  - spécialistes du TAL : des méthodes d'évaluation inspirées de la linguistique de corpus
  - lexicologues : démonstration du potentiel de l'ADA pour la description linguistique

Perspectives

- Exploitation des voisins pour réorganiser les synonymes
→ présenter en priorité à l'utilisateur les synonymes les plus pertinents pour un corpus donné

- Évaluation des ressources distributionnelles construites avec l'analyseur Talismane (Urieli et Tanguy 2013) au CLLE-ERSS
→ influence du changement d'analyseur syntaxique sur les couples extraits ?

Références

- Bloomfield, L. (1935). *Language*. Allen and Unwin, London.
- Bourigault, D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. *In Actes de TALN 2002* (24–27 juin 2002), Nancy.
- Bourigault, D. (2007). *Un analyseur syntaxique opérationnel : Syntex.* Habilitation à diriger des recherches. Université Toulouse II – Le Mirail.
- Fellbaum, C. (1995). Co-occurrence and antonymy. *International Journal of Lexicography*, 8.
- Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Norwell, MA, USA.
- Harris, Z. S. (1970). Distributional structure. *In Papers in structural and tranformational linguistics*.
- Jones, S. (2002). *Antonymy : a Corpus-Based Perspective*. Routledge Advances in Corpus Linguistics. Taylor & Francis.
- Lafourcade, M. (2007). Making people play for Lexical Acquisition with the JeuxDeMots prototype. *In SNLP'07 : 7th International Symposium on Natural Language Processing*, Pattaya, Chonburi, Thailand.

Références

- Lin, D. (1998). Automatic retrieval and clustering of similar words. In *proceedings of the 17th international conference on Computational linguistics*, pages 768–774, Morristown, NJ, USA. Association for Computational Linguistics.
- Manguin, J. L., François, J., Eufe, R., Fesenmeier, L., Ozouf, C. et Sénéchal, M. (2004). Le dictionnaire électronique des synonymes du CRISCO : un mode d'emploi à trois niveaux. In *Cahiers du CRISCO*, volume 34. CRISCO, Université de Caen.
- Morris, J. et Hirst, G. (2004). Non-classical lexical semantic relations. *In Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, CLS '04, pages 46–51, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Murphy, M. L. (2003). *Semantic Relations and the Lexicon : Antonymy, Synonymy and other Paradigms.* University Press, Cambridge.
- Urieli, A. et Tanguy, L. (2013). L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. *In Actes de TALN 2013*, (17–21 juin 2013), Les Sables-d'Olonne.
- Weeds, J. (2003). *Measures and applications of lexical distributional similarity.* Thèse de doctorat, Université du Sussex.
- Winston, M. E., Chaffin, R. et Herrmann, D. (1987). A taxonomy of part-whole relations. *Cognitive Science*, 11(4):417–444.