



# Inclusion lexicale et proximité sémantique entre termes.

Fidelia Ibekwe-Sanjuan

## ► To cite this version:

Fidelia Ibekwe-Sanjuan. Inclusion lexicale et proximité sémantique entre termes.. Terminologie et Intelligence Artificielle (TIA 2005, 2005. <sic\_00001408>

**HAL Id: sic\_00001408**

**[https://archivesic.ccsd.cnrs.fr/sic\\_00001408](https://archivesic.ccsd.cnrs.fr/sic_00001408)**

Submitted on 10 Apr 2005

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Inclusion lexicale et proximité sémantique entre termes

*Fidelia Ibekwe-SanJuan*

Université de Lyon 3

[ibekwe@univ-lyon3.fr](mailto:ibekwe@univ-lyon3.fr)

# Plan de l'exposé

- ◆ **Motivations**
- ◆ **Définitions**
- ◆ **Corpus d'étude**
- ◆ **méthodologie**
- ◆ **Règles d'ordonnancement des termes**
- ◆ **Application aux variantes "en corpus"**
- ◆ **Discussion**

# Motivations

- ◆ l'inclusion lexicale : moyen d'acquérir des relations sémantiques entre termes

- Hyperonymie / hyponymie
- Equivalence (synonymie)
- "Association" au sens large

(Morin & Jacquemin 2004 ; Bodenreider 2001 ; Grabar & Zweigenbaum 2004, etc...)

- ◆ *Pour moi* :

- ordonner les variantes des termes par proximité sémantique décroissante
- afin d'intégrer progressivement les plus proches dans une même classe

- ◆ Applications visées : VST, FT, Q-A

# Opérations d'inclusion lexicale

aka 'expansions' (Ibekwe-SanJuan, 1998) :

$t_1 = x_1 x_2$  information management

$t_2 = x_1 y_1 \dots y_n x_2$  : information **infrastructure** management

$t_3 = y_1 \dots y_n x_1 x_2$  : E-healthcare information management

$t_4 = x_1 x_2 y_1 \dots y_n$  : information management **perspective**

$t_n = y_1 \dots y_n x_1 x_2 y_1 \dots y_n$  : **personal** information management **appliance**

- ◆ Considérer toutes les positions d'ajout, quelque soit le nombre d'éléments ajoutés ou leur nature
- ◆ critères pour ordonner les variantes par proximité sémantique

# Proximité sémantique

- ◆ Une indication du degré de parenté entre le concept désigné par un terme  $t_1$  et sa variante  $t_2$  dans laquelle  $t_1$  est lexicalement inclu.
- La proximité sémantique ici  $\neq$  distance mathématique
- ◆ Pas de calcul de distance

# Proximité sémantique (2)

- Exemple d'une distance : distance d'édition
- opérations d'insertion, suppression & substitution de chaînes de caractères pour arriver de  $t_1$  à  $t_2$ 
  - ex. quel est le coût de passage de "web site" (N1 N2) à "*direct online web site*" (A1 A2 N1 N2) par rapport à "*intranet web site*" (N3 N1 N2) ou à "*web site characteristics*" (N1 N2 N3) ?
  - attribution de poids arbitraires aux opérations élémentaires
  - Pondérer ces poids éventuellement en fonction d'autres paramètres (catégorie du mot ajouté, position, ...)
  - résultats imprévisibles voire contradictoires...
- Mais possibilité : calculer similarité sémantique par ressource externe (WordNet) ==> OK pour termes de même longueur

# Proximité sémantique (3)

- ◆ Nous sommes guidés par :
  - l'introspection (ben oui !),
- + l'observation des variantes issues de corpus,
- + recherche d'indices de surface permettant de poser des hypothèses...
  - la catégorie grammaticale des éléments ajoutés (A, N, Adv),
  - la nature lexicale des noms (propre vs commun, mot-composé),
  - le nombre d'éléments ajoutés (1, 2, 3,  $n...$ ),
  - la fonction grammaticale de l'ajout (modifieur ou centre),
  - évaluer l'impact de chacun de ces indices et de leur combinaison sur la proximité sémantique entre  $t_1$  et ses variantes.



# Corpus d'étude

- **Domaine : recherche d'information (corpusIR)**
- **Taille : 3355 textes courts**
- **445 000 mots ==> base PASCAL (INIST/CNRS)**
- **49 686 candidats termes extraits**
- **3818 mots-clés (du lexique PASCAL utilisés pour indexer ces textes)**

# Méthodologie

- déterminer la relation engendrée par type d'opération d'inclusion lexicale,
- élaborer les règles de proximité sémantique en fonction des indices de surface,
- tester leur impact sur les termes extraits du corpus, considérés comme "variantes" des mots-clés de Pascal,
- plonger ensuite ces règles sur les termes du corpus uniquement.

# Relations engendrées

- ◆ **Une bijection se profile...**
  - **Ajout de modifieurs (Ins, Exp\_g)**
    - en général : hyponymie (hiérarchie)
    - mais quelques cas d'équivalence (synonymie)
  - **Ajout de centre (Exp\_d, Exp\_gd)**
    - "association" (transversale)
    - : relation "*fourre tout*"

# Relation d'hyponymie

## ◆ Ajout de modifieurs → hyponymie

1a. approximation operators :

b → *rough* approximation operators

c → \* *rough set* approximation operators

d → *pawlak* approximation operators

- 1b, c = équivalents sémantiques, “*set*” est souvent ellidé
- “*rough set*” vu comme un “bloc sémantique” (unité pré-construite)
- 1b,c,d = même niveau hiérarchique, hyponymes directs de (1a)

2a. Archival Description

b → the *Encoded* Archival Description

c → the *General International Standard* Archival Description\*

- 2b,c = = même niveau hiérarchique, hyponymes directs de (2a)

# Relation d'équivalence (?)

## ◆ Ajout de modifieurs → équivalence (?)

3a). Dewey classification → (3b) dewey decimal classification

4a). Markov model → (4b) markov chain model

- Les termes génériques (3a, 4a) :  
→ si cas d'ellipse → relation d'équivalence → même noeud
- Comment déterminer les cas de :
- hyponymie
- équivalence sémantique (ellipse) / synonymie
- si engendrées par la même opération d'inclusion lexicale ??

# Relation d'association

- ♦ Ajout de centre → association (voir aussi)

5a) african American

b → african American *households*

c → african American *low-income households*

6a) authentic reasoning → (6b) authentic reasoning *expert systems\**

7a) clustering algorithm → 7b) *robust hierarchical* clustering algorithm *ROCK\**

- Ne voit pas de critères “absolus” pour affiner cette relation...

# Implications sur la proximité sémantique

- ◆ Observations et hypothèses :
  - La compositionnalité des éléments ajoutés **diminue** la proximité sémantique (exemples 5b,c)
  - La non-compositionnalité des éléments ajoutés **accrôit** la proximité sémantique (exemples 1c, 2c, 6b, 7b)
- ◆ La thèse de la compositionnalité des énoncés :
  - "le sens d'un énoncé est l'addition des sens de ses éléments constitutifs"
- Indices de non-compositionnalité :
  - noms propres
  - noms composés
  - noms communs (hélas ! unités pré-construites : *système expert*)

# Implications pour l'ordonnancement des variantes

1. Rôle prépondérant de la “compositionnalité” des éléments ajoutés sur la proximité sémantique
2. Le rôle grammatical (centre / modifieur) des éléments ajoutés vient en 2<sup>nd</sup> position
3. Le nombre d'éléments ajoutés vient en 3<sup>rd</sup> position
  - Nos règles d'ordonnancement tiennent compte de ces constats



# Règles d'ordonnancement des variantes

<i>T1</i>	<i>Rang</i>	<i>Variante</i>	<i>T1</i>	<i>Rang</i>	<i>Variante</i>
$t_1$	exp_g1	: (<MC> N) $t_1$ : (A N)? (U NNP) $t_1$ : <SA>{1,2} $t_1$	$x_1 x_2$	ins_1	: $x_1$ (<MC> <SA> N) $x_2$ : $x_1$ (U NNP) $x_2$
	exp_g2	autres cas de type $X^+ t_1$		Ins_2	autres cas de type: $x_1 X^+ x_2$
$t_1$	exp_d3	: $t_1$ (<MC'> (A? N) : $t_1$ (A N)? NNP+			
	exp_d4	autres cas de type $t_1 X^+$			
$t_1$	exp_gd5	: (<MC'> A N) $t_1$ (<MC'> (A?N)) : (<SA> N)? NNP+ $t_1$ (<SA> N)? NNP+ NNP+ N			
	exp_gd6	: $X^+ t_1 X^+$			

# Application aux mots-clés et variantes du corpus

- ◆ Entre mots-clés de PASCAL et variantes du corpus,
- ◆ Précision théorique (intrinsèque) : validation manuelle des relations engendrées

Type d'inclusion lexicale	Nb. liens	Erreurs_Rel.	Prec._T
Expansion gauche (exp_g)	1961 (38,2)	201 (10,2%)	89,7%
Insertion (Ins)	328 (6,4%)	32 (9,7%)	90,3%
Expansion droite (exp_d)	895 (17,4%)	12 (1,3%)	98,7%
Expansion gauche-droite (exp_gd)	717 (14%)	86 (12%)	78,0%
Termes équivalents	1226 (24%)		
<i>Total Exp_g+d+gd + Ins)</i>	<b>3901</b>	<b>331 (8,5%)</b>	<b>91,5%</b>
<i>Préc. _Théorique moyenne</i>			89,1%

# Application aux variantes du corpus (2)

- ◆ 4272 termes différents du corpus impliqués dans l'inclusion lexicale
- ◆ Proportion de variantes traitées par chaque règle :

<i>Règle</i>	<i>Nb. termes</i>	<i>Nb. relations</i>
exp_g1	5241	3408
ins_1	2845	1632
exp_g2	1508	925
ins_2	606	342
exp_d3	3138	1931
exp_d4	449	258
exp_gd5	917	574
exp_gd6	246	140
<b>Total</b>	<b>14950</b>	<b>9210</b>

# Application aux variantes du corpus (2)

- ◆ Exemple d'ordonnement automatique des variantes

<i>Rang</i>	<i>Variantes</i>
$t_1$	information_NN management_NN
exp_g1	E-healthcare_NNP information_NN management_NN
exp_g1	global_JJ information_NN management_NN
ins_1	information_NN infrastructure_NN management_NN
ins_1	information_NN resource_NN management_NN
exp_g1	strategic_JJ information_NN management_NN
ins_2	information_NN security_NN risk_NN management_NN
exp_d3	information_NN management_NN perspective_NN
exp_gd5	MI5_NNP information_NN management_NN efficiency_NN
exp_gd5	labour-intensive_JJ information_NN management_NN infrastructure_NN
exp_gd5	personal_JJ information_NN management_NN appliance_NN

# Validation / Evaluation

- ◆ Vérification manuelle des 2000 premières variantes ordonnées
  - Pas d'incohérence intrinsèque entre le rang des variantes et l'hypothèse de proximité sémantique (**sommes-vous biaisés ?**)
  - Sources d'erreurs : extrinsèque aux règles
  - Erreurs d'extraction des termes en amont
    - distributed system ⇨ *\*future\_JJ* distributed\_JJ system\_NN
    - information retrieval ⇨ *\*several\_JJ* information\_NN retrieval\_NN application\_NN
    - soit 8,2% d'erreurs (164/2000)
    - 91,8% de précision théorique
- Résultat similaire à celui obtenu entre mots-clés PASCAL et variantes corpus

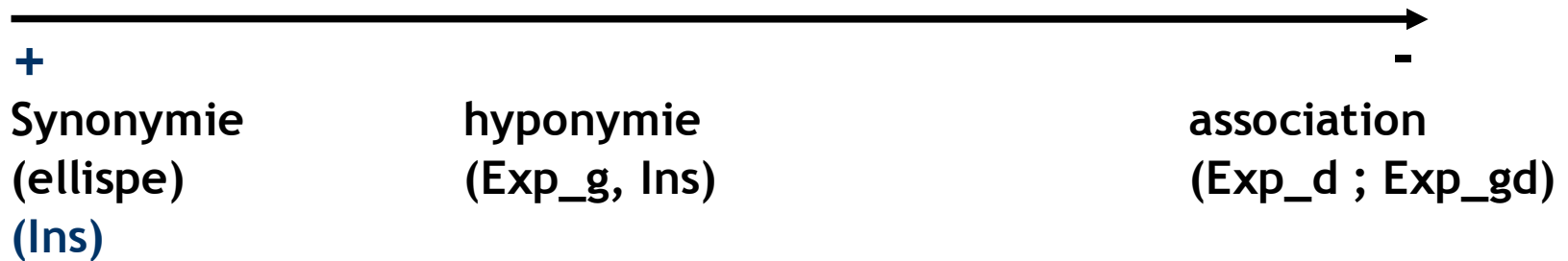
# Discussion

## ◆ Limites :

- Difficulté de détecter tous les cas de non-compositionnalité
- Des contre-exemples quand aux relations engendrées
  - ◆ (TG) algèbre de Post généralisé → (TS) algèbre de Post
  - ◆ (TG) Air bus → (TS) Air bus A320
- ◆ L'ajout de modifieurs conduit à une "généralisation" du concept lexicalement inclu (hyperonymes ?)
- Des ajouts à contenu sémantique faible
  - New York →
    - ◆ New York *City*
    - ◆ New York *State*

# Conclusions

- ◆ Ordonnancement des variantes du + proche au plus éloignée sémantiquement



- **Nécessité d'une évaluation orientée-tâche :**
  - ◆ population d'ontologies, de thesaurus
  - ◆ Q-A
  - ◆ VST, Fouille de textes



***La Fin....***

***...Merci***



# Association lexicale

- ◆ *aka* “substitutions” (Ibekwe-SanJuan, 1998)
- Pour les ordonner :
  - 1<sup>er</sup> : ressource externe (WordNet) : recherche de mêmes synsets pour mots substitués --> **Sub\_forte (relation : synonymie)**
    - comprehensive\_JJ **study\_NN**
    - comprehensive\_JJ **survey\_NN**
  - 2<sup>ème</sup> : les Sub entre termes de longueur  $\geq 3$  --> **Sub\_faible (relation : association)**
    - human\_JJ right\_NN **abuse\_NN**
    - human\_JJ right\_NN **legislation\_NN**