

MODÉLISATION LINGUISTIQUE DU CONTEXTE POUR L'EXTRACTION D'INFORMATION

Ismail El Maarouf

Lab. Valoria Université Bretagne Sud UEB

Séminaire du CENTAL - 28/10/2011

INTRODUCTION

- Extraction d'Information
 - Permettre l'accès au contenu des documents
 - Compréhension partielle
 - Recherche ciblée
 - Besoins constants
 - Quantité de textes
 - Variété des sources d'information

PLAN

1. Extraction d'information
2. Contexte et Données
3. Étude de cas
4. Modèle de segmentation
5. Relation Inter-segment
6. Relation Intra-segment
7. Sémantique et Genre

1. Extraction d'information

1. EXTRACTION D'INFORMATION

- MUC-3(1991)
 - Information structurée dans un Template

Champ	Exemple de fiche
MESSAGE: ID	DEV-MUC3-0718 (UNL/USL, NCCOSC, GE)
MESSAGE: TEMPLATE	1
INCIDENT: DATE	- 13 NOV 89
INCIDENT: LOCATION	EL SALVADOR: SAN SALVADOR (DEPARTMENT) : SOYAPANGO (CITY): PRADOS DE VENECIA (NEIGHBORHOOD)
INCIDENT: TYPE	BOMBING
INCIDENT: STAGE OF EXECUTION	ACCOMPLISHED
INCIDENT: INSTRUMENT ID	-
INCIDENT: INSTRUMENT TYPE	BOMB: "*"
PERP: INCIDENT CATEGORY	STATE-SPONSORED VIOLENCE
PERP: INDIVIDUAL ID	-
PERP: ORGANIZATION ID	"AIR FORCE"
PERP: ORGANIZATION CONFIDENCE	REPORTED AS FACT: "AIR FORCE"
PHYS TGT: ID	"HOUSES"
PHYS TGT: TYPE	CIVILIAN RESIDENCE: "HOUSES"
PHYS TGT: NUMBER	PLURAL: "HOUSES"
PHYS TGT: FOREIGN NATION	-
PHYS TGT: EFFECT OF INCIDENT	DESTROYED: "HOUSES"
PHYS TGT: TOTAL NUMBER	-
HUM TGT: NAME	-
HUM TGT: DESCRIPTION	"INJURED" "PERSONS"
HUM TGT: TYPE	CIVILIAN: "INJURED" CIVILIAN: "PERSONS"
HUM TGT: NUMBER	12: "INJURED" 3: "PERSONS"
HUM TGT: FOREIGN NATION	-
HUM TGT: EFFECT OF INCIDENT	DEATH: "PERSONS" INJURY: "INJURED"
HUM TGT: TOTAL NUMBER	-

1. EXTRACTION D'INFORMATION

- MUC-6 (1995)
 - 4 tâches : EN, TE Co-référence et Template
 - Entité Nommée

(1) Mr. <ENAMEX TYPE="PERSON"> Dooner </ENAMEX> met with <ENAMEX TYPE="PERSON"> Martin Puris </ENAMEX>, president and chief executive officer of <ENAMEX TYPE="ORGANIZATION"> Ammirati & Puris </ENAMEX>, about <ENAMEX TYPE="ORGANIZATION"> McCann </ENAMEX>'s acquiring the agency with billings of <NUMEX TYPE="MONEY"> \$400 million </NUMEX>, but nothing has materialized.

- ACE (2002)
 - Relations entre EN (Role, Part, At, Near, Social)

1. EXTRACTION D'INFORMATION

- MUC
 - Corpus spécifiques (domaine, style)
 - Protocole d'Évaluation (train/test)
 - Normalisation des mesures des systèmes

$$\text{Précision} = \frac{\text{Nombre d'éléments correctement identifiés}}{\text{Nombre d'éléments correctement identifiés} + \text{Nombre d'éléments mal identifiés}}$$

$$\text{Rappel} = \frac{\text{Nombre d'éléments correctement identifiés}}{\text{Nombre d'éléments total à identifier}}$$

$$F\text{-mesure} = \frac{(1 + \alpha) \cdot \text{Précision} \cdot \text{Rappel}}{\alpha \cdot \text{Précision} + 1 \cdot \text{Rappel}} \quad \text{avec } \alpha > 0$$

1. EXTRACTION D'INFORMATION

- Systèmes de RCEN
 - Symboliques (Rosset et al. 2006, ...)
 - Ressources et patrons
 - Supervisés (Bikel et al. 1999, ...)
 - Features et modèles statistiques
 - Semi-supervisés (Riloff et Jones 1999, ...)
 - Mots-germes et Bootstrapping

1. EXTRACTION D'INFORMATION

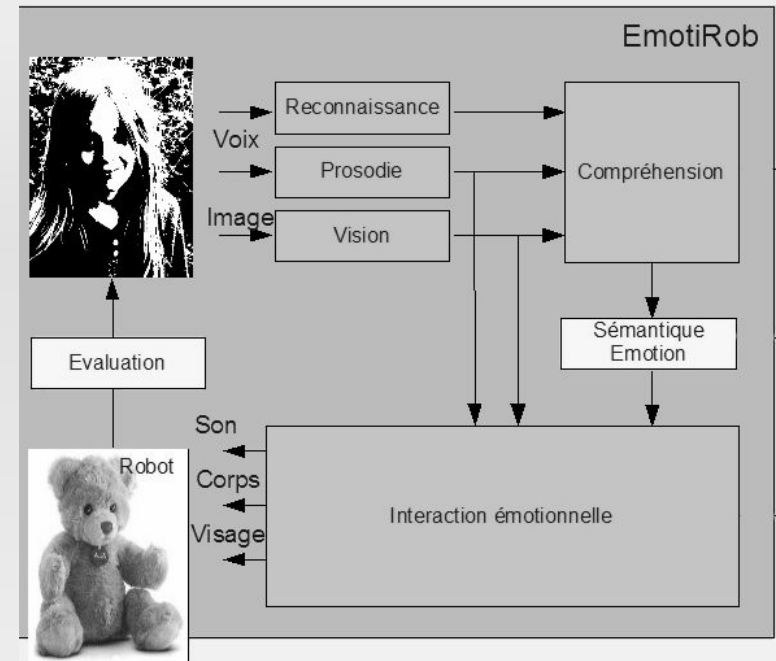
- Variation des EN
 - Forme : Majuscule, expression définies, etc.
 - Sens : Triade P-O-L, Valeurs, Produits, etc.
 - Granularité
 - + Discours : Métonymie (Semeval-7)
 - + Domaine : médical / financier, etc.
- Besoin d'analyse du contexte

2. Contexte de recherche et Corpus

CONTEXTE DE RECHERCHE

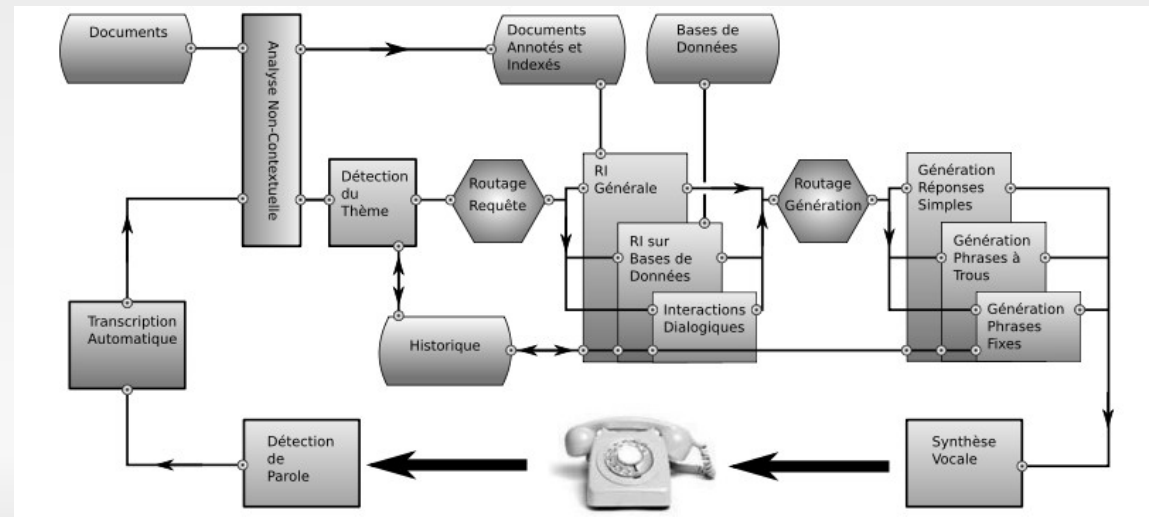
Projet EmotiRob

(ANR, Valoria, LI, Adicore)



Projet RITEL

(Limsi)



CORPUS

- Conte (EmotiRob)
 - 138 textes / 160 000 occ.
- Biographie (RITEL)
 - 430 textes / 727 000 occ.
- Presse (RITEL)
 - 43 450 textes / 22 000 000 occ.

ANNOTATION DE CORPUS

- Conte
 - Tree-tagger
 - Annotation référentielle manuelle
 - Annotation syntaxique en dépendance manuelle
- Biographie et Presse
 - Ritel-nca (entités : EN + POS + ...)
 - Chunking (Villaneau et al., 2007)
 - Segmentation discursive de surface

OBJECTIFS

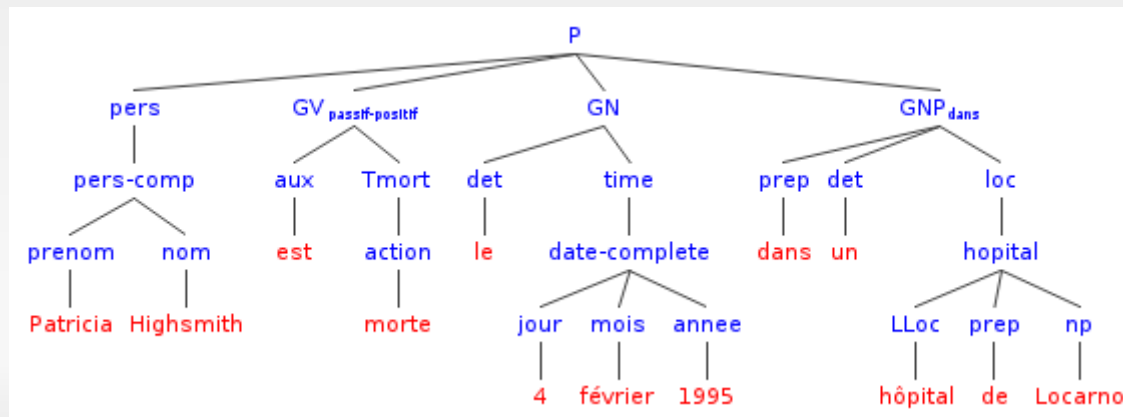
- Extraction automatique de relation sémantique
 - Appliquer des modèles sémantiques en corpus
 - Acquérir des connaissances à partir de corpus
 - Adapter les systèmes symboliques

3. Étude de cas

ÉTUDE DE CAS

- Extraction d'Information Biographique
 - Détecter les relations liées aux personnes (EN-QR)
 - En structurant leur contexte (patron lexico-syntaxique)
 - À partir d'une représentation riche

(2) Patricia Highsmith est morte le 4 février 1995 dans un hôpital de Locarno



MÉTHODE

- Linguistique de Corpus
 - Faire émerger les relations sémantiques & guider l'analyse
 - Mot-cible : unité lexicale (porteur du frame) fortement associée à l'EN
 - Environnement : Rnc + Chunking + Position
 - Profil contextuel (Smadja, 1993) d'entités

Contexte	Fréq.(y)	Fréq.(x,y)	L	R	L5	L4	L3	L2	L1	R1	R2	R3	R4	R5
<_punct>	102850	1239	498	620	0,1	0,11	0,11	0,11	0,07	0,15	0,02	0,13	0,1	0,1
<_prep>	88840	1232	488	631	0,09	0,08	0,09	0,21	0,01	0,25	0,04	0,09	0,08	0,05
<_det>	83551	1168	560	495	0,1	0,08	0,06	0,03	0,31	0,14	0,1	0,06	0,06	0,07
<_time>	11099	342	123	197	0,06	0,12	0,09	0,15	0	0	0,38	0,04	0,09	0,06
<_action>	55024	320	137	142	0,13	0,18	0,13	0,11	0,02	0,03	0,05	0,14	0,1	0,13
<_subs>	55328	310	116	154	0,13	0,14	0,17	0,05	0,01	0	0,08	0,19	0,1	0,13
<_pers>	34536	300	112	148	0,13	0,08	0,09	0,13	0,08	0,03	0,25	0,07	0,09	0,06
<_conjc>	20320	281	180	80	0,07	0,06	0,05	0,09	0,44	0,08	0,02	0,05	0,07	0,06
<_loc>	9351	256	121	101	0,13	0,13	0,18	0,13	0,02	0	0,14	0,06	0,14	0,05
<_pronom>	33791	220	100	85	0,16	0,08	0,06	0,12	0,19	0,05	0,12	0,04	0,09	0,09
<_aux>	16989	193	89	81	0,12	0,08	0,06	0,04	0,29	0,06	0,06	0,05	0,15	0,11

LIMITES

- Variation syntaxique de structures sémantiques
 - Ordre linéaire : multiplication des règles

[Pers] est né à [Lieu] en [Tps] → Naissance (Personne, Lieu, Date)
[Pers] est né en [Tps] à [Lieu]

- Interruption : Complexité des règles

(3) Raymond Lulle (Ramon Llull en catalan) (né v. 1235 à Palma de Majorque mort en 1315) était un laïc proche des franciscains (peut-être appartint-il au Tiers Ordre des Mineurs), philosophe, alchimiste, poète, mystique et missionnaire catalan du XIIIe siècle, descendant d' une famille noble catalane.

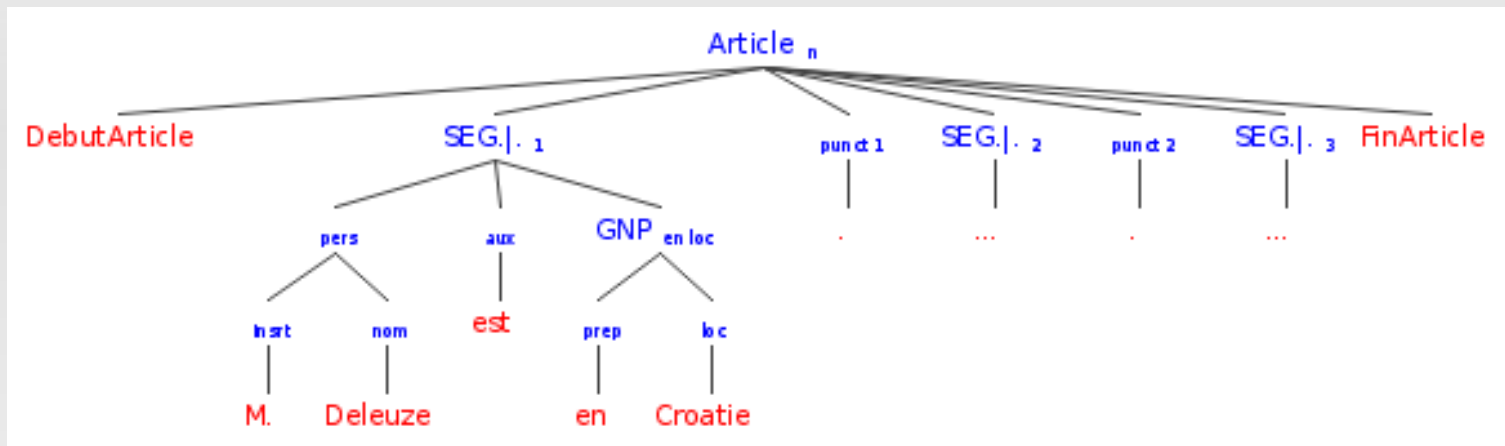
4. Modèle de segmentation

MODÈLE PROPOSÉ

- Segmentation discursive de surface
 - Segment : unité \geq chunk
 - Type de segment et classe de frontière
 - Frontière forte !?;... \rightarrow Segment fort
 - Frontière semi-faible () [] "" \rightarrow Segment semi-faible
 - Frontière faible , conjonctions et connecteurs
 - Raisonner sur les segments
 - Taille
 - Nature des frontières
 - Contenu

VISUALISER LES SEGMENTS

(4) M. Deleuze est en croatie



- Relations intra-segment (rouge) et inter-segment (violet)

A. V. Shinde, né à Goa, en Inde, décédé en 2003 à New York (à l'âge de 86 ans), avait parcouru le monde en quête des plus belles pierres d'Orient et d'Occident pour le joaillier H Winston

QUELQUES STATISTIQUES (1)

- Presse (dev-17) : 70-80 segments/article
- Nature des frontières

Formes	Fréq.	Prop.
.	1533241	6,81%
de	982808	4,37%
.	909902	4,04%
la	508511	2,26%
"	408633	1,82%
le	365872	1,63%
à	358376	1,59%
et	329005	1,46%
les	306591	1,36%
des	301536	1,34%
en	240818	1,07%
du	225005	1,00%
un	163862	0,73%
a	157251	0,70%
qui	145753	0,65%
que	144193	0,64%
dans	136519	0,61%
)	136202	0,61%
(136040	0,60%
une	135492	0,60%
pour	129899	0,58%

Type	Nb S	Nb. S cumul.	Prop. S.	Prop. S. cumul.	Nb. E	Nb. E. cumul.	Prop. E.	Prop. E. cumul.
. .	459702	459702	19,05%	19,05%	1163024	1163024	15,43%	15,43%
. .	281657	741359	11,67%	30,72%	1027209	2190233	13,63%	29,07%
. .	251479	992838	10,42%	41,14%	703517	2893750	9,34%	38,40%
"	101908	1094746	4,22%	45,36%	416932	3310682	5,53%	43,94%
. .	85170	1179916	3,53%	48,89%	417729	3728411	5,54%	49,48%
()	80543	1260459	3,34%	52,23%	276935	4005346	3,68%	53,15%
. qui	59735	1320194	2,48%	54,70%	150419	4155765	2,00%	55,15%
" .	57178	1377372	2,37%	57,07%	165839	4321604	2,20%	57,35%
. (51759	1429131	2,14%	59,22%	127678	4449282	1,69%	59,05%
. "	48790	1477921	2,02%	61,24%	156603	4605885	2,08%	61,12%
qui .	37844	1515765	1,57%	62,81%	127997	4733882	1,70%	62,82%
. "	37480	1553245	1,55%	64,36%	98686	4832568	1,31%	64,13%
. que	36387	1589632	1,51%	65,87%	94243	4926811	1,25%	65,38%
qui .	32734	1622366	1,36%	67,22%	129726	5056537	1,72%	67,10%
que .	28297	1650663	1,17%	68,40%	92051	5148588	1,22%	68,33%
que .	27286	1677949	1,13%	69,53%	112694	5261282	1,50%	69,82%
" .	23241	1701190	0,96%	70,49%	85873	5347155	1,14%	70,96%
. .	22624	1723814	0,94%	71,43%	55482	5402637	0,74%	71,70%
. qu'	22527	1746341	0,93%	72,36%	60411	5463048	0,80%	72,50%
. que	22381	1768722	0,93%	73,29%	70352	5533400	0,93%	73,43%

QUELQUES STATISTIQUES (2)

- Taille

Taille	Nb S.	Nb S. cumul.	Prop. S.	Prop. S. cumul.	Nb E.	Nb E. cumul.	Prop. E.	Prop. E. cumul.
1	616284	616284	25,54%	25,54%	616284	616284	8,18%	8,18%
2	571014	1187298	23,66%	49,20%	1142028	1758312	15,16%	23,33%
3	453120	1640418	18,78%	67,97%	1359360	3117672	18,04%	41,37%
4	282114	1922532	11,69%	79,66%	1128456	4246128	14,98%	56,35%
5	181529	2104061	7,52%	87,18%	907645	5153773	12,05%	68,39%
6	114299	2218360	4,74%	91,92%	685794	5839567	9,10%	77,50%
7	72901	2291261	3,02%	94,94%	510307	6349874	6,77%	84,27%
8	45601	2336862	1,89%	96,83%	364808	6714682	4,84%	89,11%
9	28833	2365695	1,19%	98,02%	259497	6974179	3,44%	92,55%
10	17778	2383473	0,74%	98,76%	177780	7151959	2,36%	94,91%

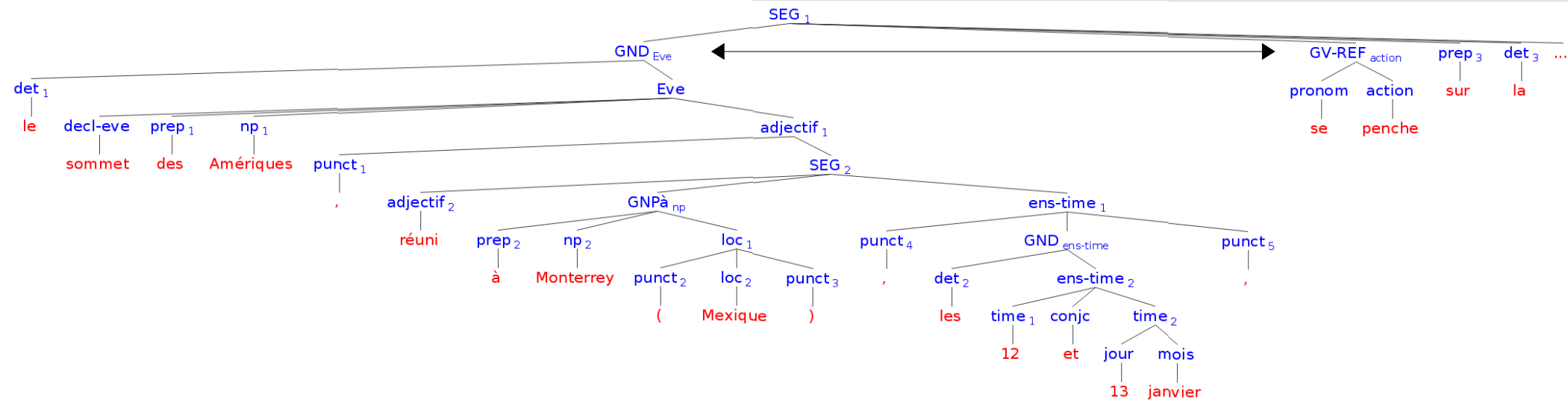
- Contenu

1-gram	Fréq.	Prop.	3-gram	Fréq.	Prop.
<i>_pers</i>	59216	7,83%	<i>_pronom _action GN_subs</i>	16045	4,34%
<i>GN_subs</i>	55565	7,35%	<i>_pronom _aux GN_subs</i>	11503	3,11%
<i>_val_score</i>	54169	7,16%	<i>_pronom _pronom _action</i>	8016	2,17%
<i>_adjectif</i>	51693	6,84%	<i>_pronom GV_action GN_subs</i>	7191	1,94%
<i>_subs</i>	42232	5,59%	<i>_action GN_subs GNP_de_subs</i>	5452	1,47%
<i>_time</i>	33863	4,48%	<i>GN_subs _action GN_subs</i>	5316	1,44%
<i>_org</i>	28987	3,83%	<i>_pronom _action _adv</i>	5072	1,37%
<i>_loc</i>	28748	3,80%	<i>_pers_bof _pronom _action</i>	4280	1,16%
<i>_action</i>	28496	3,77%	<i>_np _punct _acro_div</i>	3992	1,08%
<i>_np</i>	26824	3,55%	<i>_pronom _action _action</i>	3961	1,07%
Total	409793	54,19%	Total	70828	19,15%
2-gram	Fréq.	Prop.	4-gram	Fréq.	Prop.
<i>_pronom _action</i>	36992	6,57%	<i>_pronom _action GN_subs GNP_de_subs</i>	3147	1,23%
<i>_action GN_subs</i>	23635	4,20%	<i>_pronom _aux GN_subs GNP_de_subs</i>	2792	1,09%
<i>GN_subs GNP_de_subs</i>	17576	3,12%	<i>_pronom _action _action GN_subs</i>	2528	0,99%
<i>_pronom GV_action</i>	14370	2,55%	<i>_pronom _action _adv GN_subs</i>	1888	0,74%
<i>_action _pers</i>	13321	2,36%	<i>_pronom GV_action GN_subs GNP_de_subs</i>	1817	0,71%
<i>_action _pronom</i>	11734	2,08%	<i>GN_subs _action GN_subs GNP_de_subs</i>	1414	0,55%
<i>GN_subs _action</i>	11229	1,99%	<i>_pronom _action _pronom _action</i>	1372	0,54%
<i>_pronom GVADJ_adjectif</i>	9844	1,75%	<i>_pronom GV-NEG_action _action GN_subs</i>	1342	0,53%
<i>_prod GNP_du_time</i>	9142	1,62%	<i>_pronom _pronom _action GN_subs</i>	1280	0,50%
<i>_subs GNP_de_subs</i>	9058	1,61%	<i>_pronom _aux _adv GN_subs</i>	1274	0,50%
Total	156901	27,85%	Total	18854	7,39%

5. Relations inter-segments

RELATION INTER-SEGMENT

- Grammaire de segments (GS)
 - Parenthétiques, Insertions, Appositions, listes, etc.
- (5) Le sommet des Amériques, réuni à Monterrey (Mexique), les 12 et 13 janvier, se penche sur la ...



COMPARATIF AVEC/SANS GS

- Triplets syntaxiques

Patron	Sans	Avec	Taux d'augmentation
XVY	41,8%	52,0%	24,19%
ØVØ	10,2%	6,7%	-34,32%
XVØ	18,1%	18,8%	3,77%
ØVY	29,8%	22,5%	-24,44%

Gauche

X	Sans	Avec	Taux
<i>GN_pers_fonct</i>	0,0134	0,0167	24,9%
<i>_pers</i>	0,0483	0,0565	16,9%
<i>_np</i>	0,0180	0,0206	14,5%
<i>GN_np</i>	0,0128	0,0146	14,0%
<i>GN_pers_act</i>	0,0104	0,0117	12,5%
<i>_org</i>	0,0149	0,0166	11,5%
<i>GN_pers</i>	0,0137	0,0149	8,6%
<i>GN_org</i>	0,0222	0,0241	8,6%
<i>GN_subs</i>	0,3170	0,3390	6,9%
<i>_pronom</i>	0,3807	0,3221	-15,4%

Droite

Y	Sans	Avec	Taux
<i>_pronom</i>	0,0408	0,0435	6,8%
<i>_pers</i>	0,0200	0,0214	6,6%
<i>_action</i>	0,0315	0,0329	4,6%
<i>GVPde/action</i>	0,0344	0,0353	2,6%
<i>GVPà/action</i>	0,0212	0,0214	0,7%
<i>GN_subs</i>	0,2670	0,2689	0,7%
<i>GNPde/subs</i>	0,1197	0,1183	-1,1%
<i>GNPdans/subs</i>	0,0258	0,0253	-1,9%
<i>GNPà/subs</i>	0,0616	0,0603	-2,1%
<i>GNPpar/subs</i>	0,0253	0,0242	-4,1%

ÉVALUATION (1)

- Détection de sujet à longue distance
 - Verbe annoncer (5309 occ. 87%)
 - Variété de structures syntaxiques (sujet inv., ...)
 - Potentiel de 34 catégories (G&D) pour M1 et M2

(6) le groupe Publicis a annoncé, jeudi 8 janvier, la création de Publicis Events Worldwide

- Résultats
 - M1: XV
 - M2: X~~V~~Y

ÉVALUATION (2)

- M1

Mod.	Pos.	Catégorie	Total	Préc.	Préc. cum.	Rap. cum.	F-m cum.
M1	Gauche	<i>GN/ subs</i>	854	0,921	0,921	0,109	0,195
		<i>_pronom</i>	506	0,941	0,930	0,209	0,341
		<i>_pers</i>	213	1	0,942	0,256	0,402
		<i>_org</i>	172	0,959	0,944	0,292	0,446
		<i>GN/ Org</i>	170	1	0,951	0,330	0,490
		<i>GN/ org</i>	157	0,974	0,953	0,364	0,526
		<i>pers fonct</i>	151	0,993	0,956	0,396	0,560
		<i>GN/ pers</i>	131	1	0,959	0,424	0,588
		<i>_pers bof</i>	75	0,240	0,932	0,428	0,587
		<i>np</i>	66	0,919	0,932	0,441	0,599
		<i>GN/ org prob</i>	59	1	0,934	0,454	0,611
		<i>GN/ np</i>	50	1	0,935	0,465	0,621
		<i>GN/ fp</i>	42	0,976	0,936	0,474	0,629
		<i>GN/ pers act</i>	29	0,966	0,936	0,480	0,635
		<i>GN/ acro div</i>	30	1	0,937	0,487	0,641
		<i>prod</i>	24	1	0,938	0,492	0,645
		<i>acro div</i>	19	1	0,938	0,496	0,649
		<i>loc</i>	13	0,917	0,938	0,499	0,651
	Droite	<i>pers fonct</i>	61	1	0,940	0,512	0,663
		<i>_pers</i>	49	0,980	0,941	0,523	0,672
		<i>GN/ Org</i>	22	1	0,941	0,528	0,676
		<i>GN/ np</i>	23	1	0,942	0,533	0,680
		<i>GN/ pers</i>	20	1	0,942	0,537	0,684
		<i>loc</i>	17	0,588	0,940	0,539	0,685
		<i>GN/ org</i>	15	0,800	0,939	0,542	0,687
		<i>GN/ acro div</i>	13	0,769	0,938	0,544	0,689
		<i>org</i>	11	0,909	0,938	0,546	0,690
		<i>GN/ pers act</i>	11	0,909	0,938	0,548	0,692
<i>prod</i>		8	1	0,938	0,550	0,694	
<i>GN/ org prob</i>		6	1	0,938	0,552	0,695	
<i>GN/ fp</i>	5	1	0,938	0,553	0,696		
<i>np</i>	4	0,667	0,938	0,553	0,696		
<i>_pers bof</i>	3	1	0,938	0,554	0,696		
<i>acro div</i>	3	1	0,938	0,554	0,697		

ÉVALUATION (3)

- +M2

Mod.	Pos.	Catégorie	Total	Préc.	Préc. cum.	Rap. cum.	F-m cum.
M2	Gauche	<i>GN/ subs</i>	128	0,585	0,925	0,568	0,704
		<i>_pers</i>	67	0,846	0,923	0,580	0,712
		<i>pers fonct</i>	38	0,947	0,923	0,588	0,719
		<i>org</i>	25	0,840	0,922	0,593	0,722
		<i>np</i>	26	0,708	0,921	0,597	0,724
		<i>GN/ Org</i>	17	0,941	0,921	0,600	0,727
		<i>GN/ org</i>	16	0,938	0,921	0,604	0,729
		<i>GN/ pers act</i>	14	0,857	0,920	0,606	0,731
		<i>GN/ org prob</i>	13	1	0,921	0,609	0,733
		<i>GN/ np</i>	11	1	0,921	0,612	0,735
		<i>_pers bof</i>	10	0,625	0,920	0,613	0,736
		<i>loc</i>	8	0	0,918	0,613	0,735
		<i>GN/ fp</i>	6	1	0,918	0,614	0,736
		<i>GN/ pers</i>	5	1	0,918	0,615	0,737
		<i>GN/ acro div</i>	5	1	0,919	0,616	0,737
		<i>acro div</i>	5	0,750	0,918	0,617	0,738
		<i>prod</i>	4	1	0,918	0,618	0,738
		<i>pronom</i>	7	0,667	0,918	0,618	0,739
	Droite	<i>pers fonct</i>	30	0,897	0,918	0,624	0,743
		<i>GN/ org</i>	15	0,867	0,918	0,627	0,745
		<i>GN/ Org</i>	11	1	0,918	0,629	0,747
		<i>_pers</i>	10	0,889	0,918	0,631	0,748
		<i>GN/ org prob</i>	9	0,889	0,918	0,633	0,749
		<i>GN/ acro div</i>	6	0,833	0,918	0,634	0,750
		<i>org</i>	5	1	0,918	0,635	0,751
		<i>GN/ np</i>	5	1	0,918	0,636	0,751
		<i>GN/ fp</i>	5	1	0,918	0,637	0,752
		<i>GN/ pers act</i>	4	0,750	0,918	0,638	0,752
<i>np</i>		2	0,500	0,918	0,638	0,753	
<i>GN/ pers</i>		1	1	0,918	0,638	0,753	
<i>_pers bof</i>		2	1	0,918	0,638	0,753	
<i>acro div</i>		1	0	0,917	0,638	0,753	
<i>loc</i>	1	0	0,917	0,638	0,753		
<i>prod</i>	0	na	0,917	0,638	0,753		

6. Relations intra-segments

RELATIONS INTRA-SEGMENT

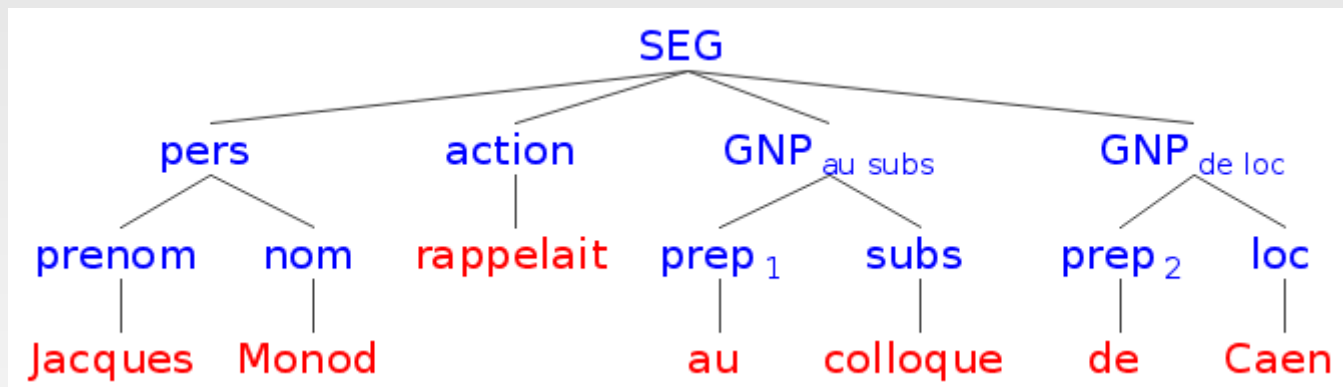
- Désambiguïsation d'EN
 - Relation sémantique et Désambiguïsation
 - Conventions d'annotation et contexte d'application
 - Extraction automatique de patron et correction

MODÈLES D'EXTRACTION (1)

- Évaluation intra-segment (>1)

- 3 modèles d'information

(7) [Jacques Monod rappelait au colloque de Caen] que ...



	<i>P1</i>	<i>P2</i>	<i>P3</i>	<i>P4</i>
<i>CF</i>	Jacques Monod	rappelait	GNP_au/colloque	GNP_de/Caen
<i>CE</i>	pers	action	GNP_au/subs	GNP_de/loc
<i>CEM</i>	pers	rappeler	GNP_au_colloque	GNP_de/loc

MODÈLES D'EXTRACTION (2)

- Modèle de score de patron

- Probabilité $PROBA(EN|Patron) = \frac{P(EN, Patron)}{P(Patron)}$

- Information Mutuelle $IM(EN, Patron) = P(X=EN, Y=Patron) \times \log \frac{P(X=EN, Y=Patron)}{P(X=EN) \times P(Y=Patron)}$

- Modèle de score de segment

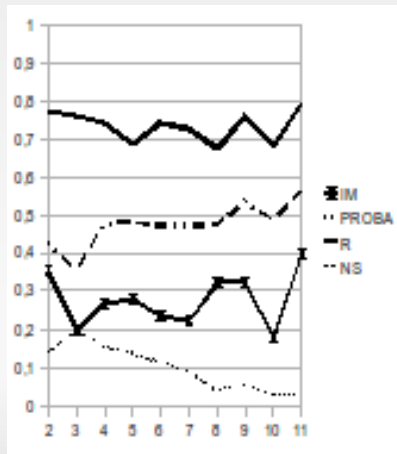
- Max, Mean, Prod $PP(Personne) = \prod_1^n PROBA(Personne|Patron_i)$

ÉVALUATION (1)

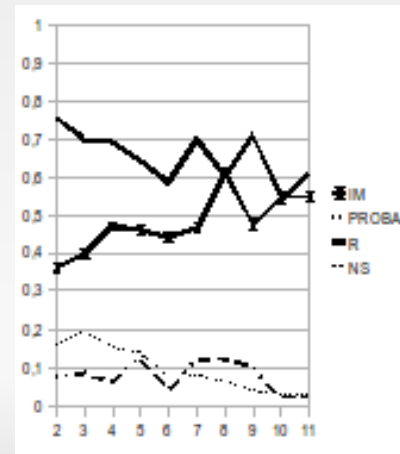
- Annotation (200 articles de presse)
 - 1426 organisations, 1004 lieux et 1377 personnes
- EN correctement détectées (Rnc)

Classe	Correct	Faux Positif	Raté	Ramenés	Annotation	PRECISION	RAPPEL	FMESURE
LIEU	689	340	135	1029	824	0,67	0,84	0,74
ORGANISATION	532	82	417	614	949	0,87	0,56	0,68
PERSONNE	1092	177	56	1269	1148	0,86	0,95	0,90

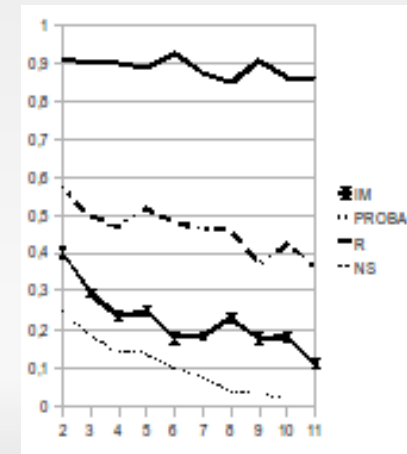
- Performance des modèles (F-mesure)



Lieu



Organisation



Personne

EVALUATION (2)

- Potentiel de correction (MAX)

Modèles	Nombre de Patrons				Nombre D'EN
	Lieu	Org	Pers	Total	
CF_PROBA_MAX	398	171	446	1015	1210
CEM_PROBA_MAX	243	112	297	652	789
CF_IM_MAX	202	74	263	539	674
CEM_IM_MAX	71	76	122	269	325
CE_PROBA_MAX	58	11	37	106	137
CE_IM_MAX	23	10	16	49	71

- Résultats

Catégorie	Modele	Precision	Rappel	F-mesure	# Corriges
LIEU	R+TOUS	0,767	0,945	0,847	91
	R+CF	0,754	0,930	0,833	78
	R+PROBA	0,726	0,939	0,819	86
	R+CE	0,727	0,910	0,808	61
	R+IM	0,739	0,886	0,806	41
	R+CEM	0,724	0,899	0,802	52
	R	0,670	0,836	0,744	NA
ORGANISATION	R+TOUS	0,952	0,686	0,797	121
	R+CF	0,941	0,672	0,784	107
	R+IM	0,917	0,666	0,772	101
	R+CE	0,925	0,624	0,745	61
	R+CEM	0,918	0,623	0,742	59
	R+PROBA	0,940	0,606	0,737	45
	R	0,866	0,561	0,681	NA
PERSONNE	R+TOUS	0,924	0,978	0,950	31
	R+CF	0,916	0,975	0,944	27
	R+PROBA	0,909	0,977	0,942	30
	R+CE	0,898	0,970	0,932	21
	R+IM	0,897	0,967	0,930	18
	R+CEM	0,894	0,969	0,930	20
	R	0,861	0,951	0,904	NA

7. Sémantique et Genre

TEXTE ET SENS

- EN ↔ Catégories ontologiques
- Type (LG, CPA) et rôle (F-Net) sémantiques
 - Ontologie (Animé, Humain, Objet, etc.)
 - Cadre (Agent, Conducteur, Vendeur, etc.)
- Restrictions de sélection
 - Type sémantique désambiguïsant
 - Pertinence pour les systèmes ?
 - Validité en corpus ?

SÉMANTIQUE DU CONTE

- Annotation référentielle

"Vous [homme] avez probablement mangé autant de lapins dans votre [homme] vie que moi [animal] et chassé autant de bêtes que nous [animal] tous ensemble."

- Alternance de type et transferts sémantiques

- Sujets du verbe *dire*

Type	Fréq.	Prop.
Humain	446	62%
Animal	107	15%
Imaginaire	98	13%
Autre	71	10%

- Conversions majeures de type

Type converti	Type de destination
Végétal, Jouet	Animé
Végétal, Animal, Jouet	Humain
Vaisselle, Animal, Végétal	Nourriture
Animé	Son

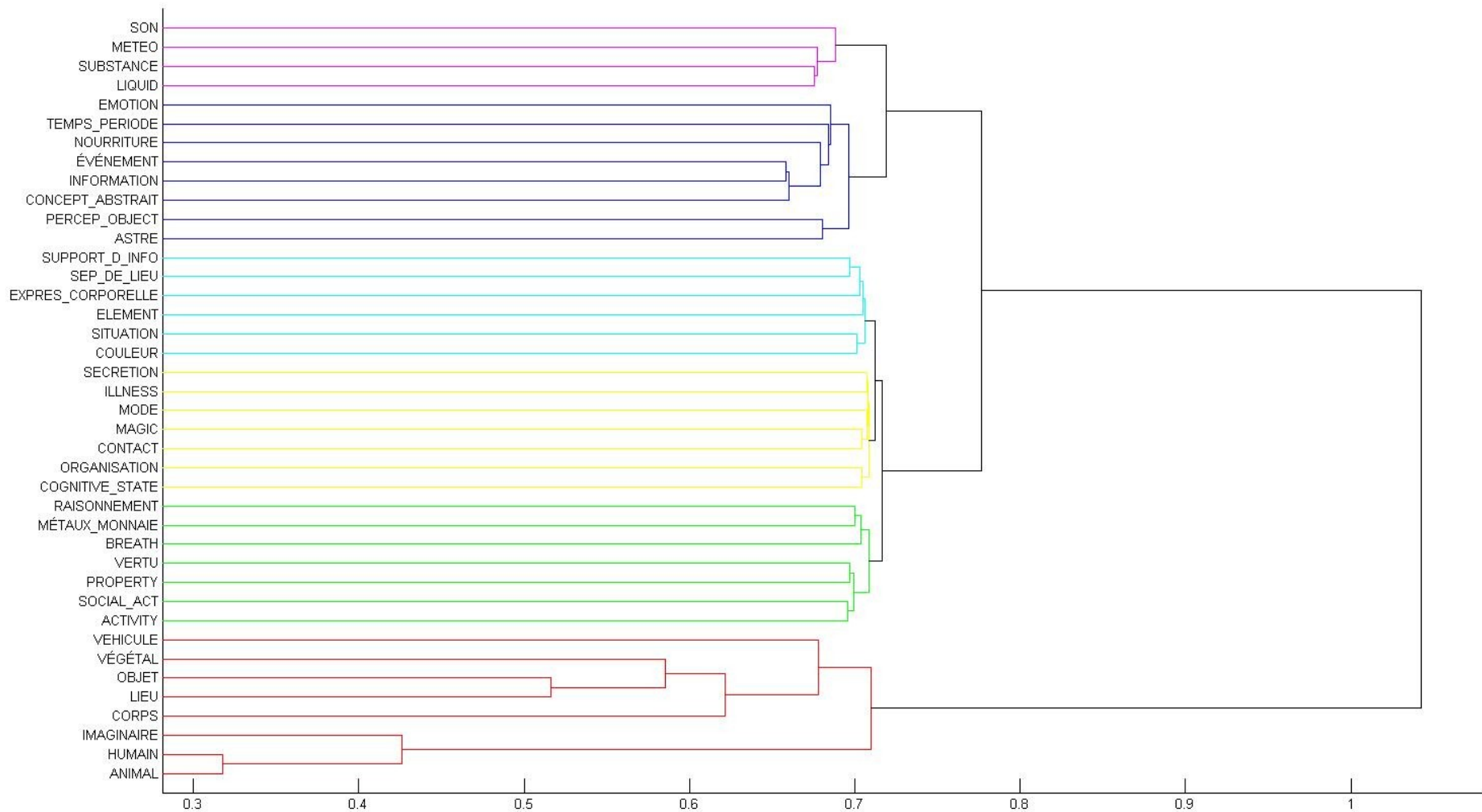
ALTERNANCE SÉMANTIQUE ET GENRE

- Outil quantitatif de caractérisation
- Couples syntaxiques + Type sémantique
- Comparaison Conte et Presse
 - Lexique commun
 - Mesure de distance (productivité) $d_{i,j} = 1 - \frac{n_{i,j}}{n}$
 - Clustering (CAH)
 - Indice de Ward (Minimiser l'inertie intra-classe)

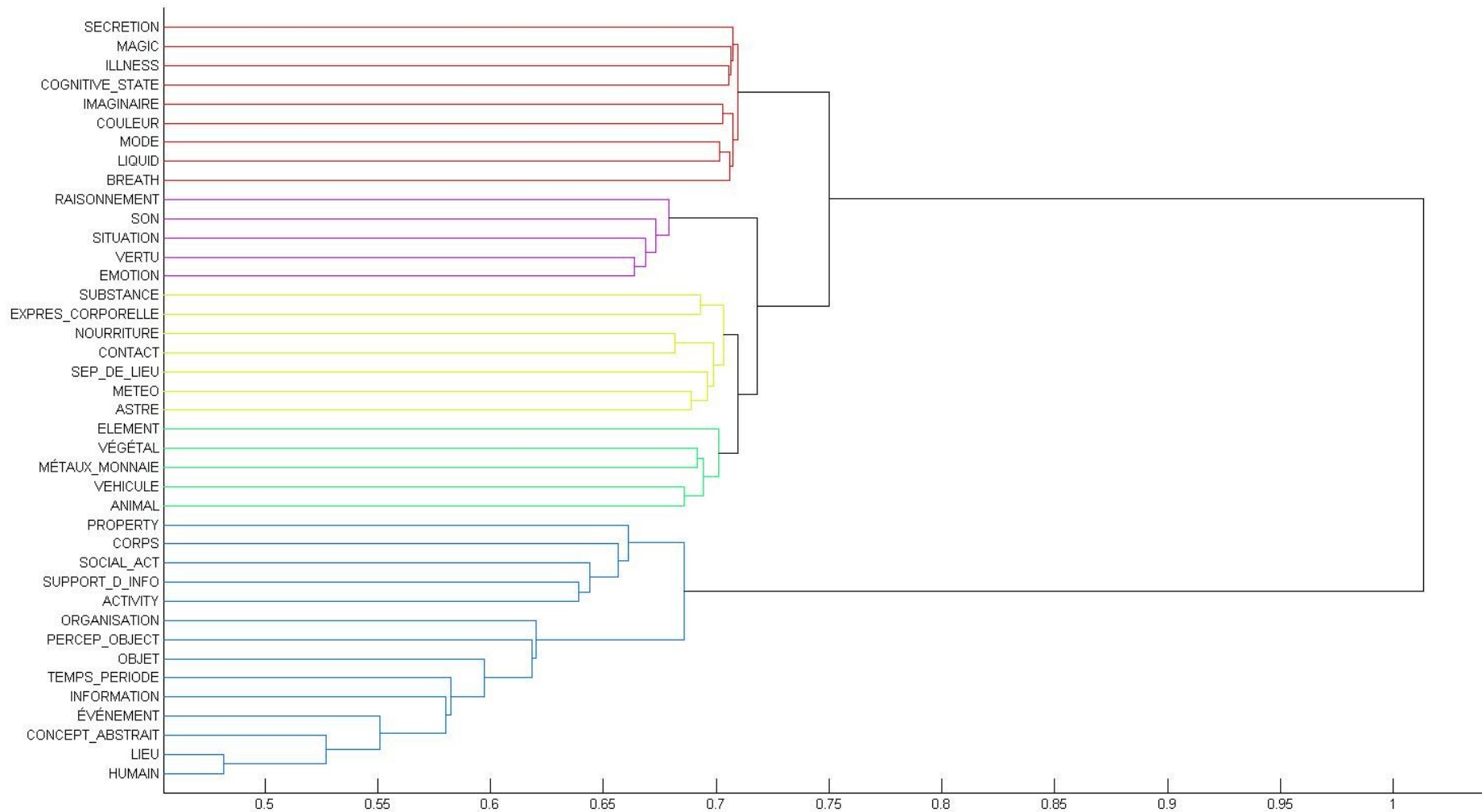
ONTOLOGIE

TYPE SEMANTIQUE	N	FT	FA	FE	PA	PE	Exemples
LIEU	208	2547	1587	960	189	275	<i>aéroport, caveme, restaurant</i>
OBJET	176	1346	445	901	125	252	<i>chaussure, bague, corde</i>
HUMAIN	151	9254	1531	7723	223	347	<i>docteur, homme, enfant</i>
ÉVÈNEMENT	92	933	712	221	140	91	<i>randonnée, vol, accident</i>
CONCEPT_ABSTRAIT	69	1252	1086	166	157	85	<i>plan, idée, intention</i>
ANIMAL	66	2996	70	2926	53	314	<i>cheval, oiseau, chien</i>
PROPERTY	51	166	121	45	71	27	<i>magnifique, lent, sage</i>
EMOTION	47	221	132	89	62	52	<i>dégoût, sensation, plaisir</i>
INFORMATION	45	893	701	192	123	60	<i>histoire, chanson, texte</i>
TEMPS_PÉRIODE	44	869	689	180	131	73	<i>année, hier, début</i>
CORPS	44	585	232	353	83	147	<i>bouche, jambe, cheveu</i>
PERCEPTUAL_OBJECT	40	439	347	92	99	58	<i>figure, carne, espace</i>
NOURRITURE	34	282	58	224	35	63	<i>viande, friandise, gâteau</i>
VERTU	31	146	90	56	47	37	<i>fierté, patience, ruse</i>
VEGETAL	30	432	56	376	35	159	<i>fleur, sapin, plante</i>
SOCIAL_ACT	21	526	465	61	83	27	<i>aide, punition, accord</i>
VEHICULE	20	208	70	138	39	79	<i>ambulance, bateau, traineau</i>
SON	19	304	149	155	56	46	<i>bruit, cri, fracas</i>
ACTIVITY	17	282	224	58	74	26	<i>escalade, jeu, lecture</i>
SITUATION	16	159	129	30	48	21	<i>repos, danger, désordre</i>
SUBSTANCE	16	121	52	69	27	54	<i>lave, sang, venin</i>
SUPPORT_D_INFORMATION	15	221	174	47	86	30	<i>livre, affiche, télévision</i>
IMAGINAIRE	14	1483	26	1457	19	226	<i>fée, sorcier, monstre</i>
METEO	14	166	63	103	41	61	<i>pluie, orage, neige</i>
ORGANISATION	13	348	313	35	112	30	<i>police, entreprise, tribu</i>
SEPARATEUR_DE_LIEU	12	190	82	108	29	40	<i>porte, paroi, barrage</i>
CONTACT	9	133	101	32	43	15	<i>baiser, fessée, choc</i>
COULEUR	9	46	30	16	19	11	<i>rouge, noir, bleu</i>
RAISONNEMENT	8	136	121	14	47	12	<i>cause, condition, explication</i>
COGNITIVE_STATE	8	24	13	11	8	11	<i>attention, folie, humeur</i>
ASTRE	6	157	32	125	29	71	<i>lune, étoile, planète</i>
EXPRESSION_CORPORELLE	6	77	54	23	29	19	<i>sourire, démarche, geste</i>
MÉTAUX_MONNAIE	6	67	55	12	33	11	<i>or, argent, monnaie</i>
MAGIC	4	25	10	15	8	9	<i>magie, sort, miracle</i>
LIQUID	3	111	31	80	14	56	<i>alcool, eau, huile</i>
BREATH	3	30	15	15	6	9	<i>respiration, souffle, soupir</i>
ILLNESS	3	11	8	3	8	2	<i>maladie, épidémie, rhume</i>
ELEMENT	2	51	19	32	15	22	<i>air, nuage</i>
MODE	2	34	29	5	10	4	<i>accent, ton</i>
SECRETION	2	4	1	3	1	3	<i>larme, sueur</i>
TOTAUX	1376	27274	10123	17151	2457	2935	

DENDROGRAMME DES CONTEES



DENDROGRAMME DE PRESSE



CONCLUSION

- Impact du genre
- Importance du contexte d'application
- Modèles d'extraction
- Méthode de segmentation

PERSPECTIVES

- Explorer les divergences entre type et rôle
- Identifier d'autres types de relations ou contextes discriminants
 - Modéliser et acquérir automatiquement des phénomènes spécifiques
- Approfondir et évaluer le modèle de segmentation (étude contrastive)

QUELQUES RÉFÉRENCES

- Bikel Daniel M., Schwartz Richard & Weischedel Ralph M., 1999, « An Algorithm that Learns What's in a Name », *Machine Learning*, vol. 34, no. 1-3, p. 211–231.
- Fillmore Charles J., , 1982, « Frame Semantics », *Linguistics in the Morning Calm*, SICOL 1981, 1982, Séoul, The Linguistic Society of Korea, p. 111-137.
- Hanks Patrick W., 2008, « Lexical Patterns: From Hornby to Hunston and Beyond », *Euralex*, 2008, Barcelone, p. 89-129.
- Nadeau David & Sekine Satoshi, 2007, « A survey of named entity recognition and classification », *Lingvisticae Investigationes*, vol. 30, no. 1, p. 3–26.
- Riloff Ellen & Jones Rosie, 1999, « Learning dictionaries for information extraction by multi-level bootstrapping », *16th national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence*, 1999, Menlo Park (Ca) (AAAI '99/IAAI '99), p. 474–479.
- Rosset Sophie, Galibert Olivier, Illouz Gabriel & Max Aurélien, 2005, « Interaction et recherche d'information : le projet RITEL », *TAL*, vol. 46, no. 3, p. 155-179.
- Sinclair John McH, 1991, *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

Merci