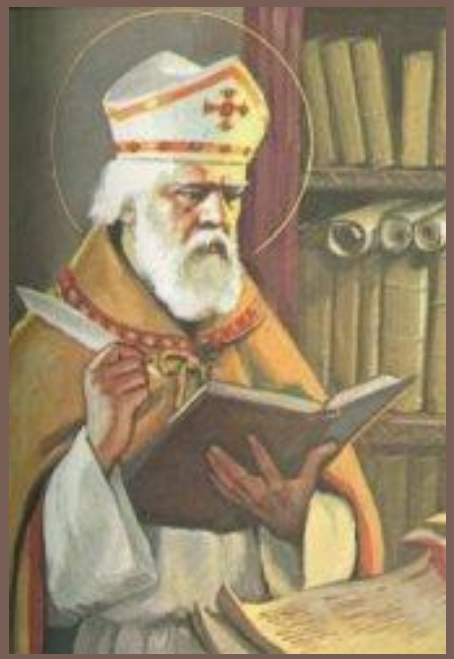


WORDNET ET SON ÉCOSYSTÈME

UN ENSEMBLE DE RESSOURCES LINGUISTIQUES DE LARGE COUVERTURE

PROXEM

FRANÇOIS-RÉGIS CHAUMARTIN

23 AVRIL 2007

Agenda

Introduction

WordNet

WordNets pour d'autres langues que l'anglais

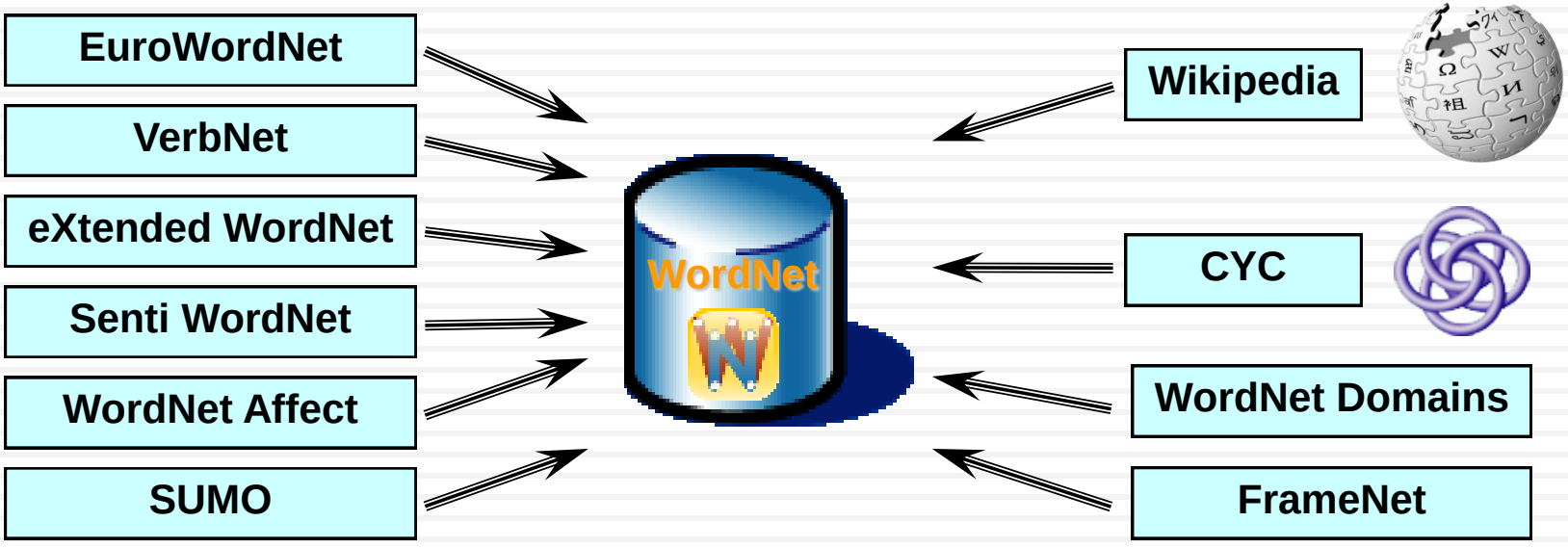
Ressources autour de WordNet

Conclusion

Questions & Réponses

Introduction

- WordNet
  - Vous le connaissez tous
  - En connaissez-vous tout ?
- Son « écosystème »



WordNet

Présentation

Synsets

Relations

Limitations

WordNet

- Base de données lexicale développée depuis 1985
  - Linguistes du laboratoire des sciences cognitives de l'université de Princeton
  - Réseau sémantique de large couverture pour l'anglais
  - Théorie psychologique du langage
  - Contenu sémantique et lexical de la langue anglaise
- L'une des ressources de TAL les plus populaires
  - Utilisable librement
  - API pour plusieurs langages

Notion de Synset

- *Synset* = ensemble de synonymes
  - Composante atomique de WordNet
  - Groupe de mots interchangeables, dénotant un sens ou un usage particulier
  - Défini d'une façon différentielle par les relations qu'il entretient avec les sens voisins
- Version 3.0 (janvier 2007)
  - 117 597 synsets
  - 207 016 lemmes
- Mappage entre les différentes versions

Organisation des Synsets

- Noms et verbes
  - Hiérarchie - Des relations d'hyperonymie et d'hyponymie relient les « ancêtres » avec leurs « spécialisations »
- Adjectifs
  - Un sens « tête » joue un rôle d'attracteur
  - Des adjectifs « satellites » lui sont reliés par des relations de synonymie
- Adverbes
  - Le plus souvent définis par les adjectifs dont ils dérivent
  - Héritent de la structure des adjectifs

Relations sémantiques (entre synsets)

Relation	Entre	Nombre	Exemple
	Verbe / verbe	13 124	EXHALE / BREATHE
Hypernym/Hyponym	Nom / nom	75 134	CAT / FELINE
Instance Hyponym	Nom / nom	8 515	EIFFEL TOWER / TOWER
Part	Nom / nom	8 874	FRANCE / EUROPE
Member	Nom / nom	12 262	FRANCE / EUROPEAN UNION
Substance	Nom / nom	793	SERUM / BLOOD
Attribute	Adjectif / nom	643	INACCURATE / ACCURACY
Verb Group	Verbe / verbe	1 748	GELATINIZE#1 / GELATINIZE#2
Verb Entailment	Verbe / verbe	409	DREAM / SLEEP
Verb Cause	Verbe / verbe	219	ANESTHETIZE / SLEEP
Adjective Similar	Adjectif / adjectif	22 622	DYING / MORIBUND
	Nom / adjectif	1 108	COMPUTER SCIENCE / ADDRESSABLE
	Nom / nom	4 146	COMPUTER SCIENCE / COMPUTER
	Nom / adverbe	37	
Topic Domain	Nom / verbe	1 236	COMPUTER SCIENCE / CASCADE
	Nom / adjectif	75	
Region Domain	Nom / nom	1 246	FRENCH / FRANCE
	Nom / adjectif	227	
	Nom / nom	563	NEUTRALIZATION / EUPHEMISM
	Nom / adverbe	73	
Usage Domain	Nom / verbe	14	
See Also	Adjectif / adjectif	2 683	BLACK / DARK

Relations lexicales (entre lemmes)

Relation	Entre…	…et	Nombre	Exemple
Usage Domain	nom	nom	379	
See Also	verbe	verbe	582	SLEEP LATE / SLEEP
Adjective Participle	adjectif	verbe	124	APPLIED / APPLY
	adjectif	adjectif	4 080	GOOD / BAD
	adverbe	adverbe	718	POORLY / WELL
	nom	nom	2 142	WINNER / LOOSER
Antonym	verbe	verbe	1 089	DIE / BE BORN
	adjectif	nom	4 814	ACADEMIC / ACADEMIA
	adverbe	adjectif	3 213	BOASTFULLY / BOASTFUL
Pertainym	adjectif	adjectif	38	
	nom	verbe	21 579	KILLING / KILL
	adjectif	nom	11 401	DARK / DARKNESS
	nom	nom	2 931	AUTOMOBILE / AUTOMOBILIST
Derivation	verbe	adjectif	1 508	KILL / KILLABLE
Adjective Cluster	adjectif	adjectif	1 290	STRIDENT / NOISY

Fréquence / Mesure de similarité

- Fréquence d'apparition pour chaque lemme
  - Indique combien de fois un mot apparaît dans un sens
  - Pour un nom ou un verbe, la somme cumulée des fréquences d'un synset et de ses hyponymes au sein d'un sous-arbre de la hiérarchie permet de calculer son Contenu Informationnel
- Métriques de « distance sémantique » entre synsets
  - Basées sur la distance à parcourir dans le graphe
  - Combinées ou non avec le Contenu Informationnel
  - Permet de quantifier la similarité de deux concepts
  - Peut servir pour la désambiguïsation lexicale
  - Exemple d'implémentation en Perl : WordNet::Similarity

Limitations de WordNet

- Structuration perfectible
- Informations manquantes
- Profusion de sens pour un mot donné
- Absence de relations pragmatiques
- Peu de corpus étiquetés par rapport à WN
  - Corpus SemCor : 676 546 mots (hors ponctuations)
    - 234 135 noms, verbes, adjectifs et adverbes ont fait l'objet d'une désambiguïsation lexicale manuelle sur WN 1.6
    - Mappage automatique vers les versions suivantes de WN

WordNets pour d'autres langues que l'anglais

EuroWordNet

BalkaNet

EuroWordNet / BalkaNet

- BD pour plusieurs langues européennes
- Phase initiale du projet achevée en 1999
  - Conception de la base de données
  - Définition de types de relations
  - Définition d'un haut d'ontologie (63 éléments)
  - Définition d'un Index-Inter-Langues (basé sur WN 1.5)
  - Néerlandais, italien, espagnol, allemand, français, tchèque, estonien
- EuroWordNet n'est pas distribué librement
- BalkaNet prolonge EuroWordNet

EuroWordNet / BalkaNet

Langue	Synsets	Lemmes	Relations internes à une langue	Relations d'équivalence entre langues différentes
WordNet 1.5	94 515	187 602	211 375	0
Ajouts à l'anglais	16 361	40 588	42 140	0
néerlandais	44 015	70 201	111 639	53 448
espagnol	23 370	50 526	55 163	21 236
italien	40 428	48 499	117 068	71 789
allemand	15 132	20 453	34 818	16 347
français	22 745	32 809	49 494	22 730
tchèque	12 824	19 949	26 259	12 824
estonien	7 678	13 839	16 318	9 004
BalkaNet				
bulgare	21 441	44 956		
tchèque	28 456	43 918		
grec	18 461	24 366		
roumain	19 839	33 690		
turc	14 626	20 310		
serbe	8 059	13 295		

Ressources liées à WordNet

VerbNet

FrameNet

eXtended WordNet

WordNet Domains / WordNet-Affect

SentiWordNet

SUMO / CYC

Wikipedia

VerbNet

- Lexique des classes de verbes anglais
  - Regroupe par classe les verbes partageant les mêmes comportements syntaxiques et sémantiques
  - Prolongement des travaux de Beth Levin
  - Identification des rôles thématiques avec d'éventuelles contraintes de sélection
  - Description de plusieurs constructions typiques (« *frames* ») des verbes membres
  - Sémantique de l'action ou de l'événement précisée
  - Sous-classes décrivent les spécialisations d'une classe
  - 237 classes de verbes regroupent 4991 sens de verbes

FrameNet

- Projet mené à Berkeley, fondé sur la sémantique des cadres (“*frame semantics*”)
- Objectif : documenter la combinatoire syntaxique et sémantique pour chacun des sens d’une entrée lexicale
  - Annotation manuelle d’exemples choisis dans des corpus sur des critères de représentativité lexicographique
  - Synthèse dans des tables qui résument pour chaque mot les cadres avec leurs actants sémantiques et arguments syntaxiques
- 825 cadres sémantiques, 10 000 unités lexicales, 130 000 phrases d’exemples annotées
- Totalité des outils et données distribuée librement (?)

FrameNet ("Crime_scenario")

A (putative) Crime is committed and comes to the attention of the Authorities. In response, there is a Criminal_investigation and (often) Arrest and criminal court proceedings. The Investigation, Arrest, and other parts of the Criminal_Process are pursued in order to find a Suspect (who then may enter the Criminal_process to become the Defendant) and determine if this Suspect matches the Perpetrator of the Crime, and also to determine if the Charges match the Crime. If the Suspect is deemed to have committed the Crime, then they are generally given some punishment commensurate with the Charges.

eXtended WordNet

- XWN produit (sur la base de WN 2.0)
  - une analyse syntaxique de la déf. de chaque synset
  - la désambiguïsation lexicale de chaque mot de la déf.
  - puis un passage en forme logique

Synsets (WN 2.0)	Nombre de définitions	Mots de classe ouverte	Mots mono-sémiques	Qualité gold	Qualité silver	Qualité normal
Noms	79 689	505 946	138 274	10 142	45 015	296 045
Verbes	13 508	48 200	6 903	2 212	5 193	30 813
Adjectifs	18 563	74 108	14 142	263	6 599	50 359
Adverbes	3 664	8 998	1 605	1 829	385	4 920

WordNet Domains

- Chaque synset (WN 2.0) est annoté avec au moins une étiquette de domaine (*Sport*, *Politique*, *Médecine*, *Economie*…), choisie dans un ensemble d'environ deux cents étiquettes organisées hiérarchiquement

Sens	Synset (Définition)	Domaines
#1	depository financial institution, bank, banking concern, banking company (a financial institution...)	Economy
#2	bank (sloping land ...)	Geography, Geology
#3	bank (a supply or stock held in reserve...)	Economy
#4	bank, bank building (a building...)	Architecture, Economy
#5	bank (an arrangement of similar objects...)	Factotum
#6	savings bank, coin bank, money box, bank (a container...)	Economy
#7	bank (a long ridge or pile...)	Geography, Geology
#8	bank (the funds held by a gambling house...)	Economy, Play
#9	bank, cant, camber (a slope in the turn of a road...)	Architecture
#10	bank (a flight maneuver...)	Transport

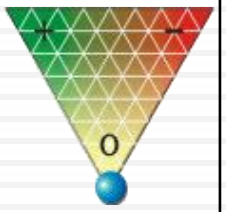
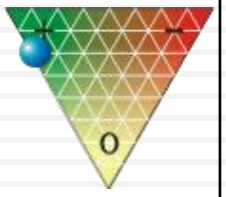
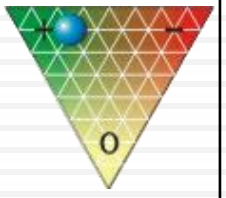
WordNet-Affect

- Ressource linguistique (basée sur WordNet Domains) pour la représentation lexicale de connaissances sur les affects

Etiquette affective	Exemples de synsets associés
Emotion	nom ANGER#1, verbe FEAR#1
Mood	nom ANIMOSITY#1, adjectif AMIABLE#1
Trait	nom AGGRESSIVENESS#1, adjectif COMPETITIVE#1
Cognitive State	nom CONFUSION#2, adjectif DAZED#2
Physical State	nom ILLNESS#1, adjectif ALL IN#1
Edonic Signal	nom HURT#3, nom SUFFERING#4
Emotion-Eliciting Situation	nom AWKWARDNESS#3, adjectif OUT OF DANGER#1
Emotional Response	nom COLD SWEAT#1, verbe TREMBLE#2
Behaviour	nom OFFENSE#1, adjectif INHIBITED#1
Attitude	nom INTOLERANCE#1, nom DEFENSIVE#1
Sensation	nom COLDNESS#1, verbe FEEL#3

SentiWordNet

- Assigne à chaque synset (WN 2.0) trois valeurs
  - Positivité, Négativité, Objectivité
  - Egalité : Positivité + Négativité + Objectivité = 1

	P = 0 N = 0 O = 1	COMPUTABLE#1 **ESTIMABLE**#3 *may be computed or estimated; "a calculable risk"; "computable odds"; "estimable assets"*
	P = 0,75 N = 0 O = 0,25	**ESTIMABLE**#1 *deserving of respect or high regard*
	P = 0,625 N = 0,25 O = 0.125	HONORABLE#5 GOOD#4 RESPECTABLE#2 **ESTIMABLE**#2 *deserving of esteem and respect; "all respectable companies give guarantees"; "ruined the family's good name"*

Suggested Upper Merged Ontology

- Une ontologie donnée est difficilement réutilisable pour une tâche autre que celle qui a motivé sa construction originelle
- Une *Upper Ontology* répertorie et organise de grandes catégories de la pensée ou de la société humaine qui devraient pouvoir être réutilisables dans de très nombreuses applications et être alors « génériques »
- SUMO est un haut d'ontologie qui se veut universel
  - Ecrit en langage SUO-KIF, dérivé simplifié de KIF
  - 20 000 termes et 60 000 axiomes

Suggested Upper Merged Ontology

- **Définition** : Any food that is ingested by drinking. Note that this class is disjoint with the other subclasses of food, i.e. meat and fruit or vegetable.
- **Sous-classes** : Milk, AlcoholicBeverage, Coffee, Tea
- **Axiomes** (traduits en anglais à partir de l'expression KIF)
  - Food is disjointly decomposed into Meat, Beverage
  - for all beverage *?BEV* holds Liquid is an attribute of *?BEV*
  - for all drinking *?DRINK* holds if *?BEV* is a patient of *?DRINK*, then *?BEV* is an instance of Beverage
  - for all Cup *?CUP* holds if contains(*?CUP*, *?STUFF*), then *?STUFF* is an instance of Beverage
  - for all Tavern *?COMPANY* holds there exist CommercialService *?SERVICE*, beverage *?BEVERAGE* so that *?SERVICE* is an agent of *?COMPANY* and *?BEVERAGE* is a patient of *?SERVICE*

Cyc

- Projet d'I.A. lancé en 1984 par Doug Lenat
  - Cyc vise à regrouper une ontologie et une base de données complètes sur le sens commun, pour permettre à des applications d'I.A. d'effectuer des raisonnements similaires à ceux des humains
  - « les chats ont quatre pattes », « Paris est la capitale de la France »…
  - Base de connaissance est divisée en plusieurs milliers de micro-théories
  - ResearchCyc compte 300 000 concepts et 3 000 000 d'assertions (faits et règles) utilisant 26 000 relations
  - Lien vers 11 300 synsets (8800 noms, 2110 verbes, 330 adjectifs et 35 adverbes) de WN 2.0

▶ Pertinent Queries (1)

All Asserted Knowledge (78)
Bookkeeping Info (1)

All KB Assertions (77)
All GAFs (50)
▼ Arg 1 (29)
▶ isa (11)
birthDate
comment
▶ conceptuallyRelated (3)
dateOfDeath
dateOfDeathEvent
definingMt
ethnicity
familyName
genStringAssertion (2)
givenNames (2)
nameString (2)
successorInPosition
synonymousExternalConcept
▶ ▼ Arg 2 (19)
conceptuallyRelated (2)
evincesBinding (3)
informationArtifactAuthor
lifetimeOf
monumentHonors
movieDirector
▶ ▼ namedAfter (2)
HistoricalPeopleDataMt
WorldGeographyMt
numberOfResultsThatSupportBindin
politicalPartyMembers

Mt : HistoricalPeopleDataMt
birthDate : (DayFn 12
(MonthFn February
(YearFn 1809)))
comment : "Abraham Lincoln (1809-1865), born in Kentucky-State, practiced law in the CityOfSpringfieldIL (Illinois-State) and held several public offices there. AbrahamLincoln was elected the 16th president of the United States and he was the Union's leader during the #$UnitedStatesCivilWar. He was assassinated by the actor JohnWilkesBooth."
conceptuallyRelated : GettysburgAddress-Speech

Mt : PeopleDataMt
conceptuallyRelated : FiveDollarBill-US PennyCoin-US

Mt : HistoricalPeopleDataMt
dateOfDeath : (DayFn 14
(MonthFn April
(YearFn 1865)))
dateOfDeathEvent : (DayFn 14
(MonthFn April
(YearFn 1865)))

Mt : BaseKB
definingMt : HistoricalPeopleDataMt

Mt : HistoricalPeopleDataMt
ethnicity : CensusGroupOfCaucasians

Mt : EnglishMt
familyName : "Lincoln"
genStringAssertion : M(nameString AbrahamLincoln "Abraham Lincoln")
M(nameString AbrahamLincoln "Abe Lincoln")
givenNames : "Abe" "Abraham"
nameString : M"Abraham Lincoln" M"Abe Lincoln"

Mt : HistoricalPeopleDataMt
(successorInPosition AbrahamLincoln JamesBuchanan President-HeadOfGovernmentOrHeadOfState UnitedStatesOfAmerica)

Mt : WordNetMappingMt
(synonymousExternalConcept AbrahamLincoln WordNet-Version2_0 "N10408858")

Wikipédia

- Encyclopédie libre et multilingue écrite de façon collaborative sur Internet avec la technologie wiki
- Plusieurs projets visent à établir automatiquement des liens entre la Wikipédia et WordNet
  - Précision de l'appariement autour de 92%

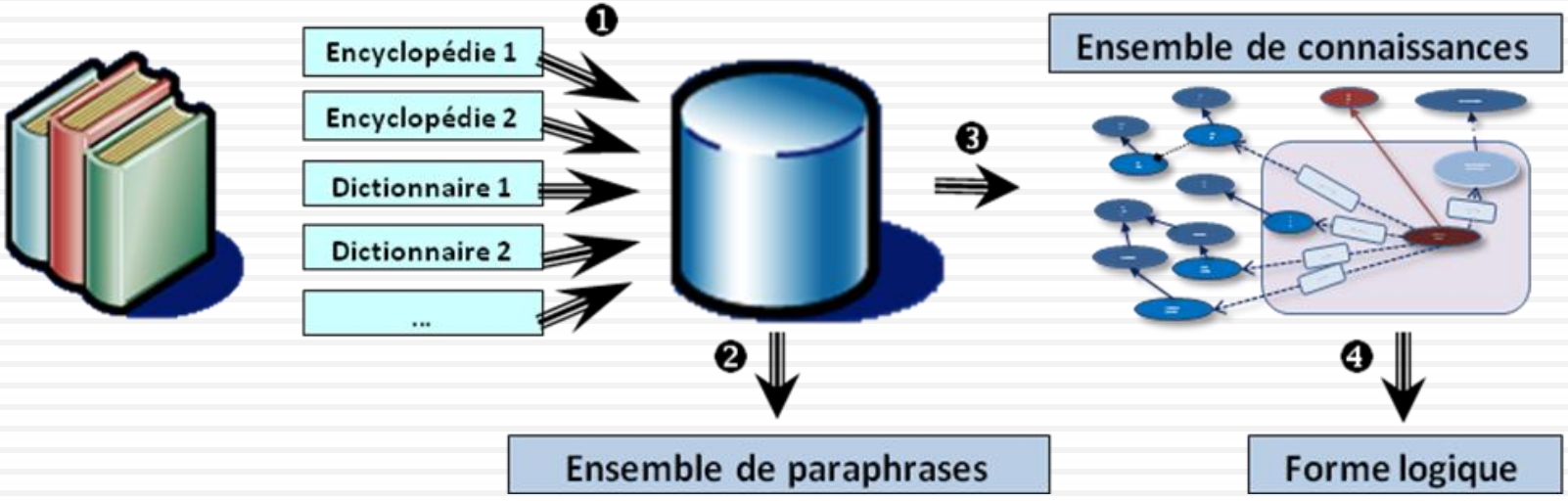
Conclusion

Bilan

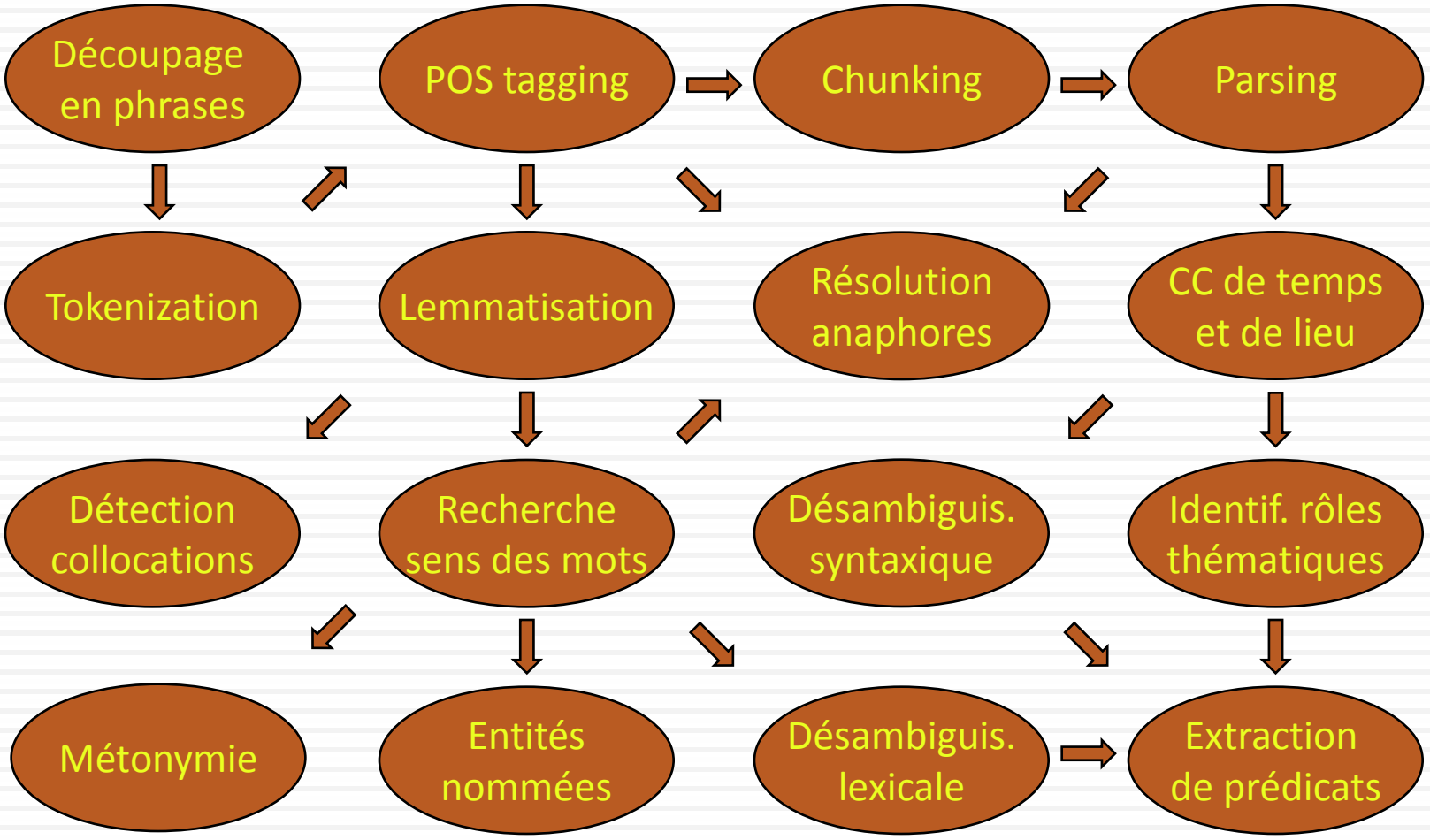
- Nous avons présenté en détail WordNet, ainsi que plusieurs autres ressources de nature lexicale, syntaxique et sémantique, qui s'y rattachent
- Le fait de mettre en commun plusieurs ressources de large couverture permet d'espérer des progrès
  - Dans les applications de TAL ou du Web sémantique
  - Recherche d'information, inférence pour la compréhension automatique de textes, désambiguïsation lexicale, résolution d'anaphores…

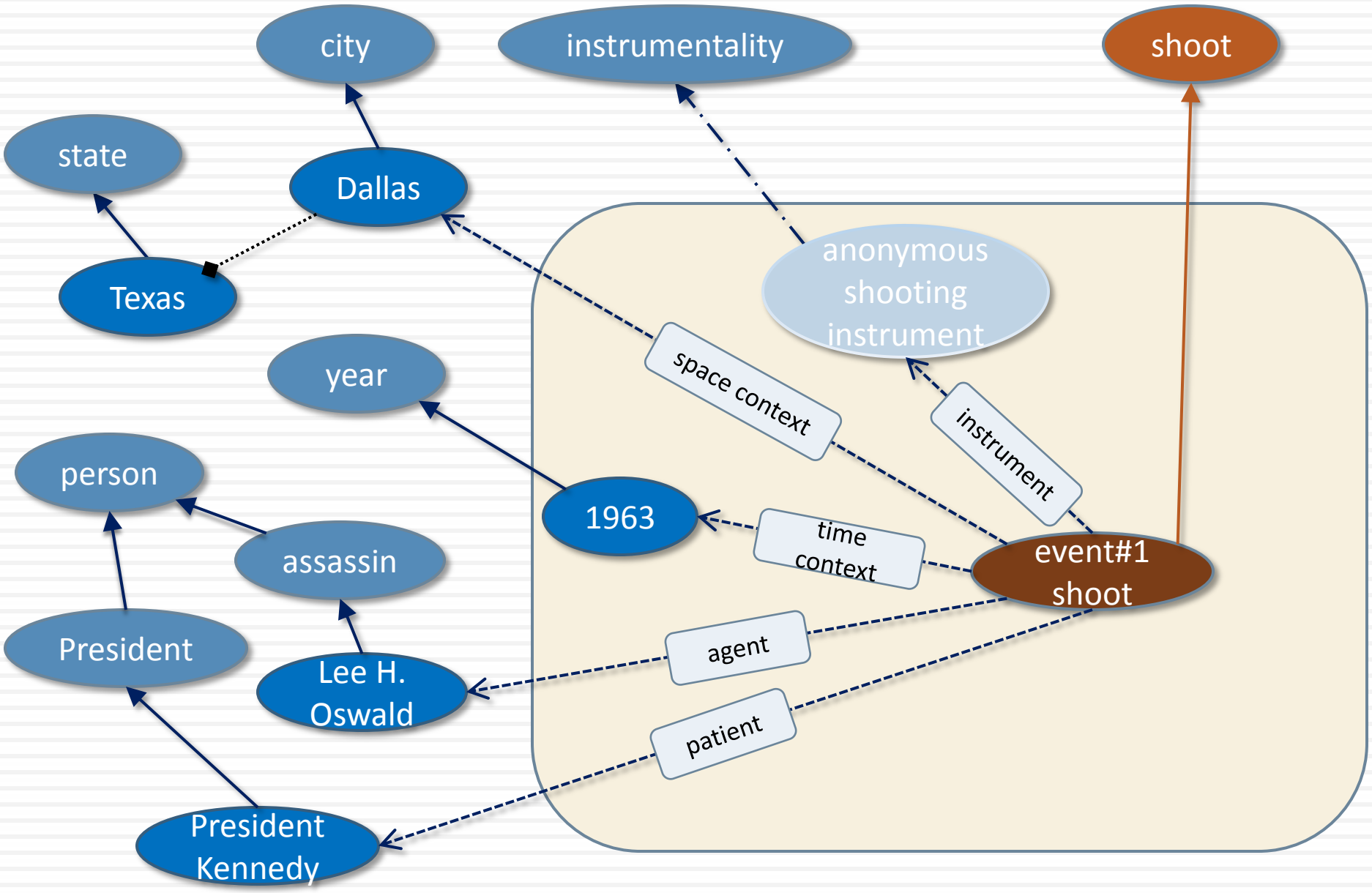
Projet ISIDORE

- Combine actuellement WordNet, VerbNet, eXtended WordNet et SUMO
  - Vise à extraire des connaissances d'une encyclopédie
  - Objectif à fin 2008 : indexation sémantique de 15 000 articles de la Wikipedia anglaise



Chaîne de traitement Antelope





"Oswald shot President Kennedy in Dallas in 1963"

Questions & Réponses

Ressources et framework .NET disponibles sur

www.proxem.com

frc@proxem.com