
Traitement Automatique des
Langues Naturelles (T.A.L.N.)
et
Construction de Ressources
Termino-Ontologiques
(R.T.O.)

Gaëlle Lortal - IF17 22/06/06

Partie I

Intelligence Artificielle

I.A.

- Avant était l'Intelligence Artificielle
 - Informatisation de processus humains
- La parole, le propre de l'homme:
 - Faire parler les machines
 - Faire agir les machines par la parole
 - Dialogue homme - machine

I.A.

- Guerre froide:
 - Traduire pour comprendre l'ennemi
 - Développement militaire
- 1950: Test de Turing (pas machine !)
 - Jeu d'imitation :
 - Imiter une conversation humaine
 - Une machine « qui pense »
 - Qui est l'ordinateur, qui est l'homme?

1950 - 1990

- Plusieurs systèmes :
 - Eliza
 - LUNAR
 - Shrdlu
- T.A.L. par les informaticiens
 - méthode des mots-clés
 - robuste et peu précise
 - méthode des automates
 - précise mais peu robuste

1990 à aujourd'hui

- Dialogue homme - machine en IA
 - Multi-agent
 - Multi-agent cognitif
- Recherche à un niveau plus profond :
 - Linguistique
 - TALN
 - Construction de ressources nécessaires

Alors...

- Informaticiens :
 - programment des formalismes universels
 - déterminent le comportement abstrait du système
- Modélisateurs :
 - produisent les ressources pour une langue et une application donnée : lexiques, expressions régulières, règles, phrases à compléter
 - déterminent le comportement des systèmes concrets

Partie II

Ingénierie linguistique

Linguistique? (1/4)

- Étude comparative et historique des langues
 - grammaire comparée, philologie comparée
 - étude en diachronie (historique) / synchronie (état)
- Science qui a pour objet l'étude du langage, envisagé comme système de signes
 - inscrite dans une visée sémiotique
 - étude des signes « linguistiques »
- Quoi étudier de la langue? la forme? le fond? Le sens?

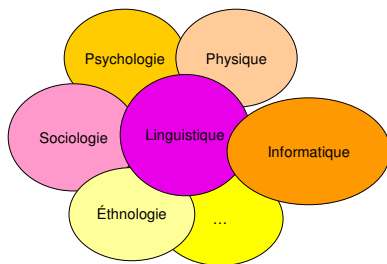
Linguistique? (2/4)

- Base de l'étude linguistique :
 - Terme
 - Phrase
 - Texte vs. Discours
 - Corpus créé / corpus réel
 - Corpus oral / corpus écrit
 - Transcription
- Différentes tendances et différents domaines

Linguistique? (3/4)

- phonétique / phonologie / ʒəmãŋzypidzavɛkdɛzoliv /
- morphologique je / mang + e/ un +e / pizza / avec /de + les / olive +s
- syntaxique [[Je] [mange [une pizza avec des olives]]]
[[Je] [mange [une pizza] [avec des olives]]]
[[Je] [mange [une pizza]] [avec des olives]]
– Je mange une pizza avec des copains
- sémantique (Je mange (une pizza avec des olives))
(Je mange (une pizza) (avec des olives))
- pragmatique (Je mange (une pizza avec des olives))

Linguistique? (4/4)



T.A.L.N. (1/3)

- Traitement automatique des langues naturelles
 - Linguistique - informatique
 - Informatique - linguistique

phonétique/
phonologie

sémantique

morphologique

pragmatique

Traduction
automatique

syntactique

T.A.L.N. (2/3)

- Informatique et modification des pratiques
 - Corpus
 - Stockage
 - Traitement
- Objectifs nouveaux
 - étude de la langue
 - ET ingénierie de la langue
 - Buts militaires (Traduction automatique, reconnaissance des langues,...)
 - Buts économiques (génération, résumé, documentation...)

Ingénierie de la langue (3/3)

- Ingénierie de la langue:
 - Système d'indexation automatique, traduction automatique / aide à la traduction, recherche d'information, système expert, documentation électronique, résumé automatique, génération de texte...
 - Automatique, semi-automatique, aide à...
- Traitement du corpus (texte) → stockage d'information → ré-utilisation

Partie III

Morphologie
Syntaxe
Sémantique

Morphologie

- Un mot ? Le système Eliza
- Morphologie flexionnelle
- Morphologie dérivationnelle
- Outils d'étiquetage

ELIZA - a friend you could never have before

Eliza: Hello. I am ELIZA. How can I help you?

ELIZA - a friend you could never have before

You: Hi! I'm Gaëlle

Eliza: Does it please you to believe I am gaëlle?

Well, I am used to

Soumettre la requête

ELIZA - a friend you could never have before

You: Well, I am used to

Eliza: How long have you been used to?

I do not know

Soumettre la requête

ELIZA - a friend you could never have before

You: I do not know

Eliza: Why do you not know?

I've never been told

Soumettre la requête

ELIZA - a friend you could never have before

You: I do not know

Eliza: Why do you not know?

I've never been told

ELIZA - a friend you could never have before

You: I've never been told

Eliza: Please go on.

I have lost my history

ELIZA - a friend you could never have before

You: I have lost my history

Eliza: Why do you say your history?

It is the best word for it

Eliza

- Méthode de compréhension de phrases sans syntaxe :
 - Énoncé est une suite de mots clés
- Analyse morphologique
 - Mots-clés reconnus ramenés à une forme standard
- Reconnaissance de patrons (patterns) composé de:
 - une liste comprenant des mots qui doivent être reconnus
 - des occurrences d'un symbole absorbant les mots non significatifs de la phrase

Eliza

- Les systèmes de dialogue basés sur les mots clé fonctionnent avec le schéma suivant :
 - Répéter
 - Afficher le message d'invite
 - Lire une phrase
 - Chercher le filtre le plus complexe reconnu par la phrase (du point de vue du nombre de mots reconnu)
 - Retourner la phrase associée au sens reconnu, en remplissant éventuellement les trous avec des parties de la phrase d'entrée capturées par le filtrage et mises sous forme appropriée (transformations morphologiques)
 - Si le sens reconnu est **stop** alors arrêter
- Système robuste, toujours une réponse

Morphologie (1/9)

- Composition des « mots » en morphèmes
 - Flexion : manger → mangera
 - Dérivation : manger → mangeoire

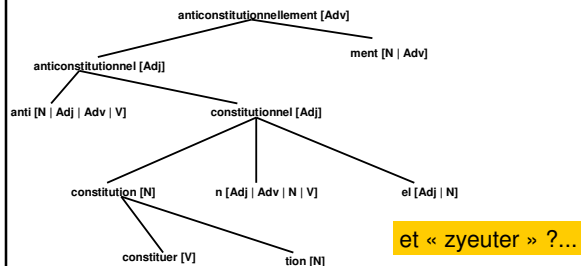
je / mang + e / un + e / pizza / avec / de + les / olive + s
je / manger / pizza / avec / olives
- Catégorisation :
 - Identifier un « mot » : pomme de terre
1.5 euros/kg
 - Analyser un « mot » : lemmatisation flexion/dérivation?
 - mangeraient → manger(aient)
 - olivier / oliveraie → olive

**collation/coller → colle*

Morphologie (2/9)

Calcul de la dérivation :

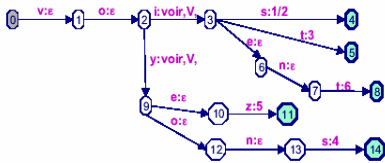
anticonstitutionnellement
 (((constituer[V]),tion[N])[N],n[Adj]Adv[N|V],el[Adj][N])[Adj],ment[N|Adv])[Adv]



Morphologie (3/9)

Calcul de la flexion : (transducteur)

Exemple : le verbe voir au présent de l'indicatif
 Etat initial 0 Etat terminal 8
 Etiquette entrée:sortie



Morphologie (4/9)

- Soubassement obligatoire du traitement du terme / du texte
- Lemmatisation :
 - étiquetage du corpus
- Lemmatisation permet :
 - un gain de stockage
 - une généralisation des termes

Morphologie (5/9)

- Méthodes à base de règles
 - Dét | Pro + Vconjugué → Pro V
 - Dét + Dét | Pro | N → Dét N
- Méthodes probabilistes
 - à partir d'un corpus d'entraînement Chaînes de Markov
 - arbres de décision binaires (TreeTagger, Schmid 1994)
- Apprentissage de règles
 - à partir d'un corpus étiqueté manuellement, induction de règles (Brill 1992)

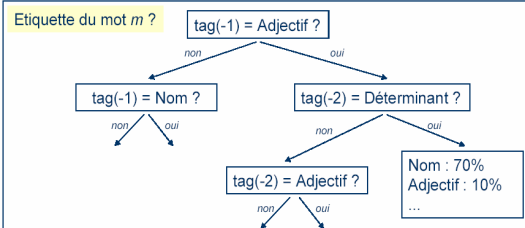
Tree Tagger

Morphologie (6/9)

- Arbres de décision binaires et estimation d'une étiquette grammaticale
- Apprentissage de trigrammes (3 étiquettes consécutives)

Tree Tagger

Morphologie (7/9)



- $P(\text{tag}_m = \text{Nom} \mid \text{tag}(-2)=\text{déterminant}, \text{tag}(-1) = \text{Adjectif}) = 70\%$
- $P(\text{tag}_m = \text{Adjectif} \mid \text{tag}(-2)=\text{déterminant}, \text{tag}(-1) = \text{Adjectif}) = 10\%$

Brill

Morphologie (8/9)

- L'étiqueteur de Brill apprend des règles d'étiquetage à partir d'un corpus annoté manuellement («Wall Street Journal»)
- A chaque étape d'apprentissage, des règles sont modifiées et le résultat de l'étiquetage avec ces nouvelles règles est comparé avec le corpus représentant l'ensemble des annotations justes
 - Utilisation des tags voisins
 - Utilisation des contextes des tags voisins
- Tant qu'un nombre d'erreurs seuil dans l'étiquetage subsiste, le processus d'apprentissage continue
- Étiquettes prédéterminées pas toujours adaptées aux textes spécialisés.

Brill

Morphologie (9/9)

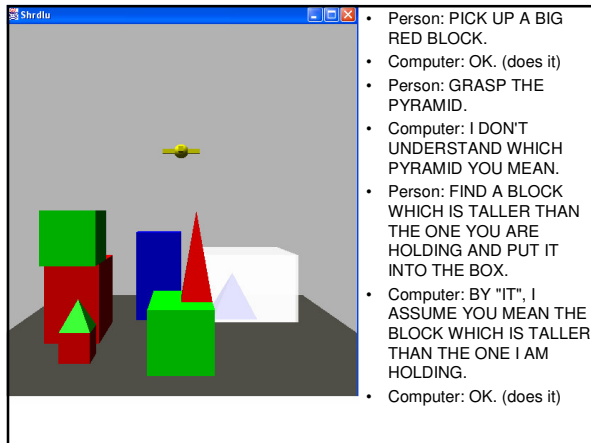
- | | |
|--|---|
| Tags (Excluding Punctuation) | Tags du Penn Tree-Bank |
| 1. CC Coordinating conjunction | 29. VBG Verb, gerund or present participle |
| 2. CD Cardinal number | 30. VBN Verb, past participle |
| 3. DT Determiner | 31. VBP Verb, non-3rd person singular present |
| 4. EX Existential "There" | 32. VBZ Verb, 3rd person singular present |
| 5. FW Foreign word | 33. WDT Wh-determiner |
| 6. IN Preposition or subordinating conjunction | 34. WP Wh-pronoun |
| 7. JJ Adjective | 35. WPS Possessive wh-pronoun |
| 8. JJR Adjective, comparative | 36. WRB Wh-adverb |
| 9. JJS Adjective, superlative | |
| 10. LS List item marker | |
| 11. MD Modal | |
| 12. NN Noun, singular or mass | |
| 13. NNS Noun, plural | |
| 14. NNP Proper noun, singular | |
| 15. NNPS Proper noun, plural | |
| 16. PDT Predeterminer | |
| 17. POS Possessive ending | |
| 18. PP Personal pronoun | |
| 19. PPS Possessive pronoun | |
| 19'. PREP Preposition | |
| 20. RB Adverb | |
| 21. RBS Adverb, comparative | |
| 22. RBS Adverb, superlative | |
| 23. RP Particle & SYM Symbol | |
| 25. TO "to" & UH Interjection | |
| 27. VB Verb, base form | |
| 28. VBD Verb, past tense | |
- Vous/PRV:pl faites/VCJ:pl preuve/SBC:sg de/PREP mesure/SBC:sg dans/PREP vos/DTN:pl propos/SBC:pl/, et/COO votre/DTN:sg discours/SBC:sgest/ECJ:sg toujours/ADV empreint/ADJ1PAR:sg de/PREP réserve/SBC:sg./, Vous/PRV:pl n'/ADV êtes/ECJ:pl certainement/ADV pas/ADV indifférent/SBC:sg./, mais/COO peu/ADV expansif/SBC:pl./, Votre/DTN:sg approche/SBC:sgplutôt/ADV formaliste/ADJ:sg peut/VCJ:sg amener/VNCF vos/DTN:pl interlocuteurs/SBC:pl à/PREP penser/VNCF que/SUB vous/PRV:pl portez/VCJ:pl une/DTN:sg grande /ADJ:sg attention/SBC:sg aux/DTC:pl conventions/SBC:pl ou/COO aux/DTC:pl usages/SBC:pl./, Votre/DTN:sg comportement/SBC:sg peut/VCJ:sg /, par/PREP contre/PREP /, paraître/VNCF assez/ADV fermé/ADJ2PAR:sg à/PREP ceux/PRO:pl qui/REL ont/ACJ:pl coutume/ADJ:sg de/PREP réagir/VNCF spontanément/ADV /, Votre/DTN:sg approche/SBC:sg sérieuse/ADJ:sg peut/VCJ:sg amener/VNCF vos/DTN:pl interlocuteurs/SBC:pl à/PREP penser/VNCF que/SUB vous/PRV:pl considérez/VCJ:pl le/DTN:sg temps/SBC:sg comme/SUB un/DTN:sg**

Syntaxe

- Un autre jeu, Shrdlu
- Formalismes syntaxiques
- Algorithme d'analyse syntaxique
- Outil d'étiquetage syntaxique

Shrdlu

- Monde des blocs
- Mise en place de dialogues finalisés avec un but
- Contrôle de la compréhension par des actions non langagières



Shrdlu

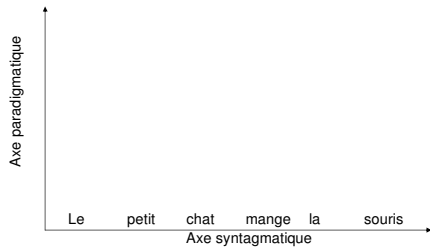
- Person: PICK UP A BIG RED BLOCK.
- Computer: OK. (does it)
- Person: GRASP THE PYRAMID.
- Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
- Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
- Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
- Computer: OK. (does it)

Shrdlu

- Modèle du monde des blocs :
 - ((SUR TABLE BLOC1) (SUR BLOC1 BLOC2) (SUR TABLE BLOC3) (ROBOT TENIR BLOC4))
- Comprend des ordres et les réalise
- Comprend des questions sur le monde ou sur l'histoire du dialogue et y répond
- Capacités de raisonnement et de planification pour obtenir les résultats demandés
- Base syntactico-sémantique

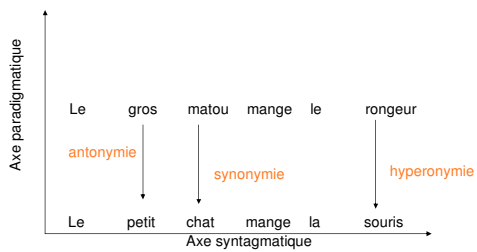
Syntaxe (1/24)

- Structure syntagmatique : catégories et frontières des constituants



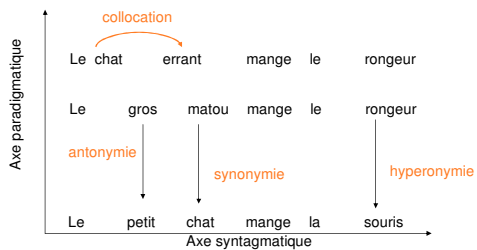
Syntaxe (1/24)

- Structure syntagmatique : catégories et frontières des constituants



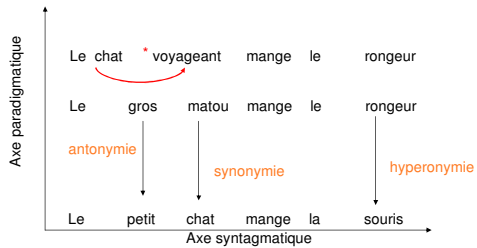
Syntaxe (1/24)

- Structure syntagmatique : catégories et frontières des constituants



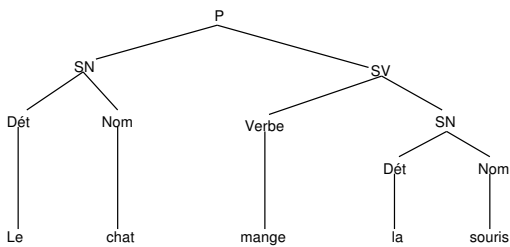
Syntaxe (1/24)

- Structure syntagmatique : catégories et frontières des constituants



Syntaxe (2/24)

- Structure syntagmatique : catégories et frontières des constituants



Syntaxe (3/24)

- Fonctions ou relations de dépendance



Syntaxe (4/24)

- Postulat de départ :
 - pas de sémantique sans syntaxe
- Syntaxe computationnelle :
 - Chomsky *Syntactic Structures* 1957
 - point de jonction entre linguistique et informatique
- Comment transposer l'analyse des langages formels aux langues naturelles ?
 - formalismes pour coder les connaissances syntaxiques (HC, GT)
 - algorithmes d'analyse syntaxique (parsing)

Syntaxe (5/24)

- Grammaires Hors Contexte (HC) :
 - Grammaire HC : ensemble de règles de réécriture de la forme
 - La grammaire HC décrit la structure syntagmatique canonique de la phrase (structure profonde)
 - $P \rightarrow SN\ V\ SN \mid SN\ V$
 - $SN \rightarrow N \mid Pro \mid Det\ N \mid SN\ que\ P$

La pizza que tu as mangée avait des olives

Et...*Je te promets de venir*
Le chat a mangé la souris et le renard le chat
La souris est mangée par le chat ???

Syntaxe (6/24)

- Grammaire transformationnelle (Chomsky):
 - Grammaire HC + règles de transformation
 - Des transformations s'appliquent à une structure profonde pour produire différentes paraphrases (structures de surface)
 - $SN(1)\ V(2)\ SN(3) \rightarrow SN(3)\ auxiliaire_{être}\ V(2)_{Part,passé}\ par\ SN(1)$
Le chat mange la souris / La souris est mangée par le chat
 - Pas de limites à l'application des règles de transformation
 - Pas de relation de dépendance
 - Génération d'une structure, mais pas d'analyse

Syntaxe (10/24)

- Structure de « chart » (Kay 1986) : on stocke dans une table tous les constituants bien formés au fur et à mesure qu'ils sont construits pour éviter de les recalculer
- Technique du coin-gauche (ou du coin-tête) : on stocke également les structures partiellement construites pour éviter de tenter de refaire des analyses qui ont échoué
- Compactage de l'analyse (Tomita 1985)
 - Empaquetage des ambiguïtés locales
 - Partage de sous-arbres
 - Résultats : forêts d'analyse

Syntaxe (11/24)

- Performances des analyseurs:
 - L'analyse est de complexité polynomiale en fonction du nombre de mots de la phrases
 - Selon l'implémentation, le nombre maximal de catégories en partie droite des règles est également un facteur à prendre en compte
 - D'autres facteurs peuvent être limitants en pratique
 - Le nombre de règles
 - Les temps d'accès au lexique
 - Des performances insuffisantes pour analyser précisément de gros volumes de textes tout venants

Syntaxe (12/24)

- Analyse syntaxique en composant (constituant et dépendances)
- Extraction de la terminologie (puis classification) pour construction de ressources termino-ontologiques
- Approches linguistiques (vs statistique | mixte)

FASTR Syntaxe (16/24)

- Problème:
 - « *build association rule* »
 - association rule
 - association and classification rules
- ➔ association de **domaine** et règles de classification

LEXTER Syntaxe (17/24)

- Extraction de groupe nominaux maximaux
- Décomposition de groupes nominaux maximaux
- Présentation des résultats sous forme d'un réseau sémantique

LEXTER Syntaxe (18/24)

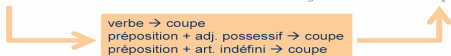
- Extraction des groupes nominaux maximaux:
 - Découpage du texte par repérage de la frontière
 - Patrons morpho-syntaxiques de marqueurs de frontière (verbe, conjonction, préposition + adj.poss,...)
 - Isolement de syntagmes nominaux susceptibles d'être des occurrences de termes

Texte initial (étiqueté)

le circuit d'aspersion de
l'enceinte de confinement
assure le maintien de sa
température nominale de
fonctionnement après une
augmentation de pression.

Groupes nominaux maximaux

- circuit d'aspersion de
l'enceinte de confinement
- maintien
- température nominale de
fonctionnement
- augmentation de pression

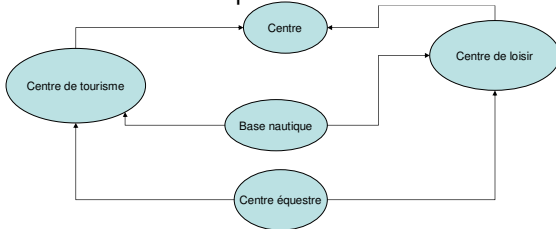


LEXTER Syntaxe (19/24)

- Décomposition des groupes nominaux maximaux
 - Terme complexe =
tête + expansion/modifieur
nom1 + adjectif
nom1 +(de)+ nom2
 - Ambiguïtés :
/centre de /tourisme/ équestre/

LEXTER Syntaxe (20/24)

- Présentation des résultats sous forme de réseau sémantique



XIP Syntaxe (21/24)

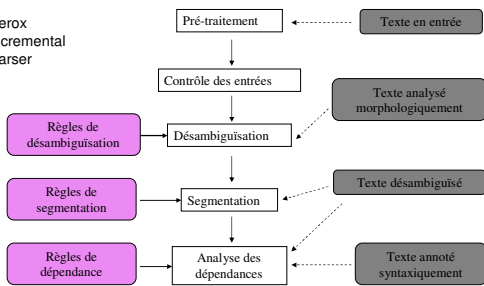
- Vue d'ensemble:
 - Analyseurs utilisant des modèles traditionnels avec un mécanisme particulier pour éliminer les échecs ou sélectionner la meilleure analyse (Frank *et al.*, 98: extension LFG)
 - Approches probabilistes : calcul de la structure la plus probable en utilisant des règles extraites d'un corpus annoté
 - Analyseurs partiels : structures minimales éventuellement sous spécifiées, néanmoins utilisables pour une analyse plus complète
 - Systèmes hybrides

XIP Syntaxe (22/24)

- sélection et marquage contextuel pour la désambiguïsation morphologique
- marquage structural (chunking) : construction d'un arbre de syntagmes noyau à partir d'une liste de mots
- calcul de dépendances : extraction de relations de dépendance à partir des têtes des syntagmes de l'arbre

XIP Syntaxe (23/ 24)

Xerox
Incremental
Parser



Syntaxe (24/24)

- Analyse/ Calcul des dépendances
 - Une dépendance est une relation n-aire qui connecte des noeuds selon une relation spécifique :
 - dépendances syntaxiques standard (sujet ou objet)
 - relations de dépendance plus complexes comme des relations entre phrases (corréférence).
 - Le calcul de dépendances prend comme entrée l'arbre de constituants et met en relation les têtes des syntagmes
Ingénierie ← ingénierie des connaissances
- Méthode structurale (structure syntaxique interne des candidats termes)

Syntaxe → sémantique

Sémantique

- Théories sémantiques
- Étiquetage sémantique

Sémantique (1/4)

- Méthode contextuelle d'extraction de dépendances:
 - Exploitation du contexte de co-occurrences des candidats-termes
 - Locales : une relation extraite pour une occurrence
 - patrons
 - Globales : relations extraites à partir d'un ensemble d'occurrences
 - cooccurrence statique
 - analyse distributionnelle

Sémantique (2/4)

- Méthode contextuelle locale, patron :
 - Un ... est un ... qui*
 - Tous les ... sauf ...*
 - ... et ...*
- Problème de généralisation des patrons, acquisition des patrons et validation.

Sémantique (3/4)

- Méthode contextuelle globale (cooccurrence) :
 - Cooccurrence statique :
 - 1er ordre : unité qui cooccurrent avec le mot pivot dans une fenêtre donnée (phrase, paragraphe)
 - 2ème ordre : unités qui ont les mêmes cooccurents que le mot pivot
 - Analyse distributionnelle:
 - 2 termes sont rapprochés s'ils apparaissent dans les mêmes contextes syntaxiques (outil Zellig)
 - Mesures de proximité,...

Sémantique (4/4)

- Récupération de termes
- Récupération de relations syntaxiques
- Inférence de relations paradigmatiques

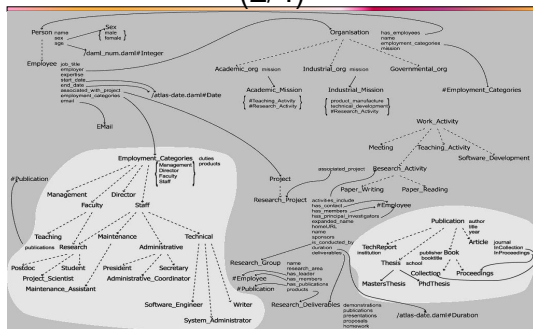
→ création d'une RTO

Utilisation de la RTO pour l'étiquetage sémantique de corpus, l'indexation pour recherche d'information, navigation, etc.

Ressource termino-ontologique (1/4)

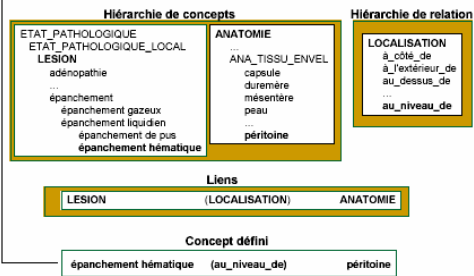
- Ressource informatique décrivant le vocabulaire et les concepts spécifiques à un domaine, à une communauté pour le traitement de l'information
- Composition d'une R.T.O. :
 - termes + relations
 - Mots / groupes de mots
 - Relations est_un, voir_aussi, relations syntagmatiques,...
- Structuration de R.T.Os :
 - index, lexique, dictionnaire, thesaurus, ontologie

Ressource termino-ontologique (2/4)



Ressource termino-ontologique (3/4)

→ hémopéritoine : « épanchement hématique localisé au niveau du péritoine »



Ressource termino-ontologique (4/4)

- Construction d'une R.T.O. :
 - Partir d'un texte
 - Par traitement statistique : algorithmes matriciels,...
 - Par traitement linguistique : extracteur de terme, extracteur de relations entre les termes, classification de termes
 - Appel à différents niveaux d'étude de la langue :
 - Morphologie / syntaxe / sémantique
- Traiter le terme et sa structure syntaxique pour extraire du « sens »
