

Outils et techniques de Traitement Automatique des Langues

Goritsa NINOVA
Tech-CICO, ICD - UTT
IFO9- Systèmes documentaires

Plan du cours

- Les raisons d'être du TAL
- Eléments de théorie linguistique
- Présentation des outils et des techniques du TAL
- Présentation d'un domaine d'application

23/05/07

Goritsa Ninova IFO9

2

I. Qu'est ce que le TAL ?

- L'objet du traitement - la langue
 - Différents types de données linguistiques
 - Texte
 - Dialogues écrits et oraux
 - Groupes de mots
 - On peut les manipuler : tester, modifier, agencer, construire ou reconstruire
- Automatique = qui opère par des moyens mécaniques
 - Capacité de ramener les manipulations sur les données linguistiques à des calculs
 - Automatisation
 - Totale
 - Partielle – assistée par ordinateur
- Traitement
 - Outils et techniques de traitement
 - Connaissances et techniques linguistiques précises
 - Procédures formelles permettant d'exprimer des connaissances en formalismes
 - Techniques informatiques

23/05/07

Linguistique et Traitement Automatique des
langues C. Fuchs, 93

3

I. Que fait le TAL ?

- Des outils
 - Correcteurs orthographiques
 - Traducteurs automatiques
 - Génération automatique de textes
- Des outils d'accès au contenu textuel
- Des ressources pour les outils
- Des formalismes de représentation des ressources

I. Aborder les aspects techniques du TAL – une double préoccupation

- Exploiter
 - Les apports de plusieurs siècles de linguistique à travers les modèles et connaissances accumulés
 - Les apports de la modélisation informatique qui assure le socle formel et technologique de l'ensemble des réalisations du domaine

I. L'ordinateur face à la langue

- Linguistique et Informatique
 - Quelle syntaxe ?
 - Quelle sémantique ?
 - Quelle pragmatique ?
- Langage, cognition et Intelligence Artificielle
 - Place centrale à la notion de connaissances

Décrire des langues naturelles

- Ambiguïté
- Variation
- ...

II. Éléments de théories linguistiques

- Les unités
 - Quelques implicites fondamentaux
- Les niveaux d'analyse

II. Éléments de théories linguistiques

Les unités

- Phonème
- Morphème
- Lexème
- Syntagme
- Phrase
- Texte

II. Éléments de théories linguistiques

Les unités et les niveaux linguistiques

- ❑ Son et phonème
 - Ce sont les phonèmes qui nous permettent de nous comprendre malgré nos accents étrangers.
 - raison vs saison vs rasons vs rayon vs raisin
- ❑ Morphèmes lexicaux, et morphèmes grammaticaux
 - *Chant - oz, fenêtre - s;*
 - *Calcul - ette, im - possible*
- ❑ Le lexique
 - Lexique général et lexique de spécialité
 - Lexique et grammaire (? *Le canard rit.*)
- ❑ Syntagme et fonction
 - *Le marié de ma voisine lave sa voiture. VS Il nettoie.*
 - *Une robe saumon*
- ❑ Phrase
 - *Tu arrives? - Oui.*
 - *Elections Législatives en Grenade-Bretagne*
 - *Il a préparé son voyage et il est parti. Il a préparé son voyage; il est parti.*
 - *Le voyage est un loisir. Comme le jeu.*
- ❑ Phrase et texte

II. Eléments de théories linguistiques

Cohésion et cohérence du texte

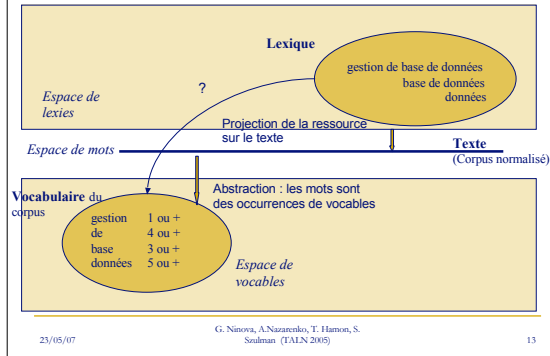
- Qu'est ce que la cohésion?
 - « *Il fallut un temps à Mégret pour mettre la main sur l'assassin du ministre. Il a cru tout d'abord... Puis il a cherché du côté de... Après bien des péripéties, il a enfin trouvé la trace de ce criminel abominable.* »
 - *Finalemnt il l'a arrêté à Lyon VS*
 - *Finalemnt il a été arrêté à Lyon*
- Qu'est ce que la cohérence?
 - Question : *Pourquoi le professeur Tournesol vient-il à la Sorbonne en patins à roulettes?*
 - Trois réponses proposées:
 - *Parce qu'il a cours.*
 - *Parce que le métro est en grève.*
 - *Parce qu'il est fou.*

II. Eléments de théories linguistiques

Quelques implicites fondamentaux

- oral/écrit
- Le mot
 - ❑ Mot graphique / mot phonétique
 - ❑ Mot sémantique / mot lexical
 - ❑ Mais quelle est donc la plus petite unité significative?
- Principe de double articulation du langage
 - ❑ Unités distinctives, non pourvues du sens : *les phonèmes*
 - ❑ Unités associant forme et sens : *les morphèmes*
- Langue/discours, sens/signification, sémantique/pragmatique
 - ❑ *Veuillez attacher vos ceintures !*
 - ❑ *Il fait froid.*
- Axe syntagmatique / axe paradigmatique

Mot, texte et ressource



III. Outils et formalismes pour le TAL

- Mots et analyse lexicale
- Phrase et analyse syntaxique
- Sens et traitements automatiques des langues
- Méthodes d'accès aux ressources linguistiques

23/05/07

Gonitsa Ninova IF09

14

III. Décomposition de l'analyse automatique en sous tâches

- Progrès de l'automatisation de la TA
 - Décomposition en sous tâches:
 - Analyse du texte source
 - Transfert
 - Génération du texte cible
 - Décomposition de la phase d'analyse
 - Analyse morphologique (identification des mots ou des morphèmes)
 - Analyse syntaxique (identification des syntagmes et de leurs fonctions)
 - Analyse sémantique

23/05/07

Gonitsa Ninova IF09

15

III.1. Mots et niveau lexical

- Enjeux applicatifs variés
 - Détection et correction d'erreurs
 - Documentation, indexation, moteur de recherche
 - Étiquetage lexical
- Objectif de l'étiquetage lexical – reconnaître et étiqueter des formes pertinentes qui ont un statut d'unité de base = auxquelles est attaché un sens
- Difficultés de l'étiquetage lexical
 - Les informations linguistiques sur les mots ne sont pas déductibles de leur forme
 - Pluriel ou singulier : *raisons vs stimulus et prends vs réseaux*
 - La taille du vocabulaire de mots simples (sens typographique)
 - Toutes les langues présentes des ambiguïtés lexicales : *règle : nom/verbe*
 - certains systèmes d'écriture n'ont pas de séparateurs
 - problème de séparation de mots *fibre optique*

23/05/07

G.Ninova IF09

16

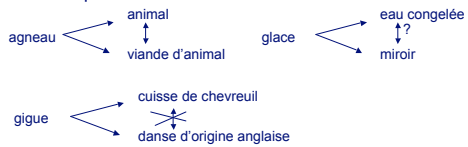
III.1. Mot et niveau lexical

L'ambiguïté lexicale des langues

Catégorielle



Sémantique



23/05/07

Gonisa Ninova IF09

17

III.1. Mots et niveau lexical

Outils actuels

- Données spécifiques du langage : dictionnaires électroniques, grammaires électroniques, collections de textes
 - Élément central dans la conception d'un traitement
 - Nécessitant la prise en compte du volume et le format des données
- Formalismes et modèles formels
- Techniques et méthodes de base

23/05/07

Gonisa Ninova IF09

18

1. Mot et niveaux lexical

Dictionnaires électroniques ou lexiques

Information linguistique associée à l'entrée

phonologique

morphologique Manger =

syntactique

manger [cat=verbe]
partir [sous-catégorisation = intransitif]
partir [sous-catégorisation = (Arg1 = X
Arg 2 = Y)

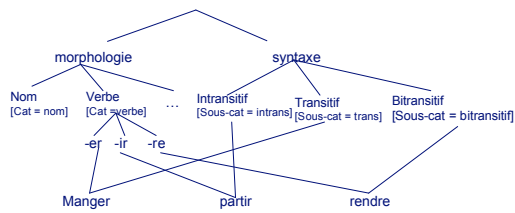
Sémantique

Je prends du sable dans la main Massique/comptable
*Je prends du castor dans la main
?Le frère est venu Prédicatif/ relationnel
Le frère de Pierre est venu
Mon frère est venu

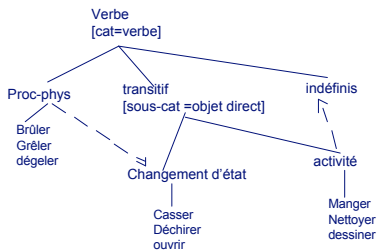
Information flexionnelle

(Nombre = pluriel
Personne = troisième
Temps = imparfait
Mode = indicatif)

Le lexique contient beaucoup d'informations
Besoin de les regrouper de façon économique



Les lexiques d'héritage - conçus comme une hiérarchie



Exemple de corpus étiqueté BulTreeBank

```
<bbt1pos>
: <S>
<S><NP>
<w aa="Noun" la="Noun">Автомобил</w>
<w aa="Conj:Noun:Prepos" la="Prepos">под</w>
<w aa="Noun" la="Noun">кваса</w>
<S></NP>
<S><VP>
<w aa="Verb" la="Verb">означав</w>
<w aa="Adj" la="Adj">превозно</w>
<w aa="Noun" la="Noun">средство</w>
<S></VP>
<w aa="Verb" la="Verb">конструирано</w>
<w aa="Conj:Part" la="Conj">и</w>
<w aa="Verb" la="Verb">оборудвано</w>
<w aa="Conj:Part" la="Conj">да</w>
<w aa="Verb" la="Verb">превоза</w>
<w aa="Noun:Prepos" la="Prepos">до</w>
<w aa="Num" la="Num">десет</w>
<w aa="Noun" la="Noun">лџици</w>
<w aa="Conj:Part" la="Conj">и</w>
<w aa="Noun" la="Noun">бгажа</w>
<w aa="Pron" la="Pron">им</w>
```

23/05/07

<http://www.bulreebank.org/>

22

III.1. Mot et niveau lexical Techniques d'analyse lexicale

- segmentation (identification des frontières des mots et des mots composés)
- lemmatisation (identification du mot sous sa forme canonique)
- Étiquetage (identification de la bonne catégorie morpho-syntaxique pour une forme donnée)

23/05/07

Gonina Ninova HF09

23

Le principe de la reconnaissance en morphologie

- On part d'une chaîne de caractères et on essaie de la découper de façon à ce qu'à chaque segment corresponde une unité répertoriée dans le système

Il a mangé des pommes de terre
10 listes possibles

Quelques exemples célèbres d'ambiguïté

Le pilote ferme la porte. Le cuisiner salue la note. La petite brise la glace.

23/05/07

Gonina Ninova HF09

24

Technique informatique de reconnaissances de formes fléchies

■ Analyse procédurale

- Automate à états finis
 - Lire les lettres une par une en partant de la dernière
 - À certains états est associé une terminaison valide du dictionnaire des flexions
 - L'analyseur regarde dans le dictionnaire des radicaux si un radical de la classe flexionnelle adéquate existe
 - Si c'est le cas – une solution est trouvée

23/05/07

Gonitsa Ninova IF09

25

Technique informatique de reconnaissances de formes fléchies

■ Approche déclarative

- Les connaissances linguistiques sont séparées de l'algorithme lui-même
- Possibilité de mise en œuvre sur plusieurs jeux de données
- On découpe les mots de toutes les manières possibles et comparant chaque fois les deux parties du mot aux données
- On génère un dictionnaire des formes fléchies et l'algorithme consiste à la recherche dans le dictionnaire la ou les forme(s) correspondante(s) au mot analysé

23/05/07

Gonitsa Ninova IF09

26

III. 2. Grammaires et analyseurs syntaxiques

- La syntaxe étudie comment la phrase se structure.
- Elle est décrite dans une grammaire
 - Définition des principes et des contraintes régissant la combinatoire des mots
 - Distinction des phrases correctes des phrases incorrectes
- La grammaire a une double fonction
 - Fonction normative
 - Règles de combinaison des mots
 - Fonction représentative
 - La grammaire associe une phrase à sa ou ses représentation(s) syntaxique(s)

23/05/07

Gonitsa Ninova IF09

27

Exemple de grammaire

- R1 $P \rightarrow SN SV$
- R2 $SN \rightarrow Npr$
- R3 $SN \rightarrow DET N$
- R4 $SN \rightarrow DET ADJ N$
- R5 $SV \rightarrow V SN$
- R6 $SV \rightarrow V$
- R7 $DET \rightarrow \{le, la\}$
- R8 $N \rightarrow \{\text{chat, pomme, lait}\}$
- R9 $Npr \rightarrow \{\text{Jean}\}$
- R10 $V \rightarrow \{\text{mange, court, boit}\}$

23/05/07

Gonita Ninova IF09

28

Les connaissances syntaxiques

- Nécessité de prise en compte des phénomènes de micro-syntaxe
 - Grammaires des dates
 - Grammaires des nombres
 - Grammaires des signes de ponctuation
 - Analyse des locutions

23/05/07

Gonita Ninova IF09

29

La représentation des connaissances syntaxiques

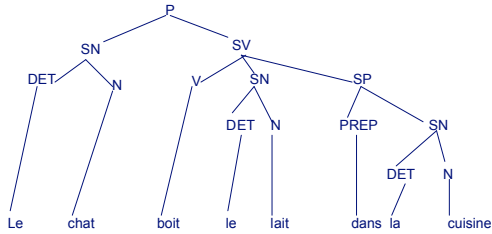
- Grammaires de constituant
 - Règles de réécriture
 - Règles lexicales
 - Structures de traits
 - Règles de précédence
 - Typage
- Grammaires de dépendances
 - Bien adaptées au TAL car lexicalisées
 - Fréquemment utilisées pour des applications utilisant l'analyse superficielle
- Grammaires d'unification
 - Décomposition des catégories en traits
 - Principe de partage de valeur de traits entre syntagmes
 - Même type de représentation pour le lexique, la syntaxe et la sémantique

23/05/07

Gonita Ninova IF09

30

Exemple de structure des constituants



23/05/07

Gonisa Ninova IF09

31

Les grammaires d'unifications

- Approche de surface
- Structure de traits
- Unification
- Stratégie inductive
- Optique déclarative

23/05/07

Gonisa Ninova IF09

32

La structure de trait

- Deux caractéristiques importantes
 - Elle est réursive
 - Elle admet le partage de valeurs



23/05/07

Gonisa Ninova IF09

33

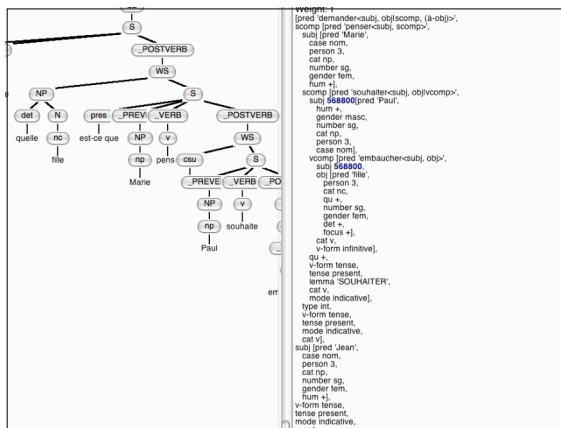
Les analyseurs syntaxiques : principes

- Représentation des connaissances syntaxiques
 - Deux sortes de connaissances syntaxiques
 - Des règles
 - Des informations attachées aux lexiques (lexique –grammaire)
 - Quantité de connaissances considérable
 - Techniques de représentation permettant de les entrer, de les modifier et de les compléter facilement
 - Décrites dans des fichiers texte dans des formats « lisibles » pour un humain qui peut les contrôler et modifier
 - Utilisées par le système d'analyse
 - Directement – interpréteur
 - Après avoir été transformées – compilateur

23/05/07

Gonissia Ninova IF09

34



Stratégie d'analyse

- Analyse descendante – dirigée par les hypothèses
- Analyse ascendante – dirigée par les données
- Avantages et inconvénients
- Stratégies
 - Choix du sens de la lecture
 - Technique des « charts »
- ...

23/05/07

Gonissia Ninova IF09

35

Traitement des ambiguïtés locales

- Stratégie de retour en arrière
 - Choix d'une solution arbitraire
 - Ordre de plausibilité, heuristique de choix
 - Mémorisation des solutions non retenues et l'état de l'analyse au moment du choix
- Stratégie du parallélisme
 - Développer toutes les solutions possible lorsque un choix se présente
 - Difficulté technique d'écriture d'algorithmes simulant le parallélisme
- Stratégie alternative
 - Analyseurs déterministes – technique de «regard en avant»

23/05/07

Gonitsa Ninova IF09

37

Les corpus syntaxiquement annotés

- L'analyse syntaxique prend en entrée une phrase et lui assigne une ou plusieurs représentations syntaxiques
- deux phases d'analyse
 - Étiquetage morpho-syntaxique (tagging)
 - Analyse syntagmatique (parsing)
 - A chaque étape une phase automatique et une phase de correction et d'enrichissement manuelles

23/05/07

Gonitsa Ninova IF09

38

III.3. Sens est TAL

- Le sens
 - Sens et référence
 - *l'étoile du soir VS l'étoile du matin*
 - Signification et vérité
 - *Marie dort.*
 - Représentation sémantique et interprétation
 - Monde réel et représentation mentale

23/05/07

Gonitsa Ninova IF09

39

Les phénomènes sémantiques

- Deux types d'indices pour le calcul du sens
 - Le sens des mots identifiés par l'analyse morphologique
 - Sémantique lexicale
 - Le sens des relations entre mots
 - Sémantique grammaticale

Sémantique lexicale

- Deux grands types de structuration des sens lexicaux utilisés en TAL
 - Décomposition sémantique
 - Réseaux sémantiques

Sémantique lexicale Décomposition sémantique

- Recours à des traits sémantiques
 - Humain, animé
 - *Un professeur de judo blond.*
- Décomposition sémantique
 - Recherche des « sèmes » = traits sémantiques minimaux, dont la composition constitue le sens des unités lexicales
 - « avoir des bras »
 - *Fauteuil vs chaise*

Sémantique lexicale

Réseaux sémantique

- Graphes
 - Nœuds qui représentent les unités
 - Arcs qui représentent les relations
- Les relations
 - *Le merle est un oiseau*
 - *Une roue est une partie d'une voiture.*
 - *Une scie sert à couper*
- Difficultés
 - Définition précise d'un lien
 - *Une chambre -?-partie d'un appartement*
 - *Une instituteur -?-partie de L'Education Nationale*
 - La nature des unités qu'on relie
 - Nombre de type de liens et maîtrise et la cohérence du système

23/05/07

Gonissia Ninova IF09

43

Sémantique grammaticale

- Traduction de la structure de la phrase en formule logico-sémantique
 - Interprétation des relations sémantiques
 - Pas de correspondance immédiate
 - Une structure - représentation sémantique différents:
 - *Jean habite à Paris.* (état sans agentivité du sujet)
 - *Jean regarde le ballon.* (activité du sujet est sans résultat)
 - *Jean attrape le ballon.* (changement de l'état de l'objet sous l'action de l'agent)
 - L'inverse
 - *Cette clé ouvre la porte.*
 - *La porte s'ouvre avec cette clé.*

23/05/07

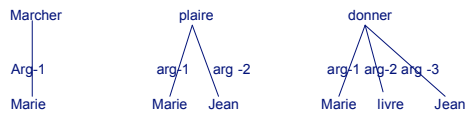
Gonissia Ninova IF09

44

Sémantique grammaticale

Interprétation des relations syntaxiques

- Représentation en termes de prédicat-argument
 - Règles d'analyse sémantique - permettent après consultation du **dictionnaire** de transformer l'arborescence syntaxique en une représentation prédictive



23/05/07

Gonissia Ninova IF09

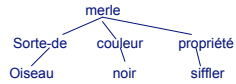
45

Les formalismes de représentation

Deux grandes familles

- Structures d'attribut-valeur
 - Représentation de la décomposition sémantique
 - Représentation des réseaux sémantiques

merle
Sorte de : oiseau
Couleur : noir
Propriété : siffler



- Formalismes logiques

- MERLE(x)
- DONNER(x,y,z)

Analyse sémantique

Principes

- Relation entre analyse syntaxique et analyse sémantique
 - L'entrée de l'analyse sémantique et l'arbre syntaxique associé à la phrase
 - Idée de correspondance « règle à règle »
 - Sémantique compositionnelle

III.4. Méthodes d'accès aux ressources linguistiques

- Concordances et statistiques lexicales élémentaires
 - Présenter une suite de lignes de contextes
- Recherche de collocations dans un texte
 - Mot pôle
 - Fenêtre d'observation
 - Test statistique
 - Mise en évidence des groupements lexicaux dont la cooccurrence est significative

Définition d'un marqueur (Ségèla, 2001)

- **Notion de marqueur**- formule linguistique qui atteste de façon plus ou moins stable l'expression d'une relation sémantique entre termes. A un marqueur est associé un schéma – le résultat de l'encodage dans un langage opérationnel de sa formule linguistique
- **Langage de définition des schémas:**
 - **Une relation** : HYPONYMIE
 - **Un identifiant de schéma** : Y_EST_LE_X_LE_PLUS
 - **Un schéma permettant d'extraire la relation**
Hyponymie(X,Y) : Y ETRE (1*PLUS) ART_DEF X ART_DEF

23/05/07

Gonisa Ninova IF09

52

Classification de termes

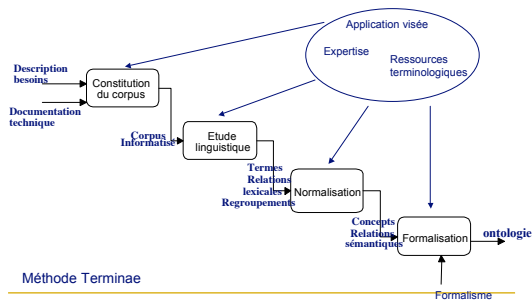
- Classique : mots associés
 - regroupement de mots apparaissant dans les mêmes contextes textuels (fenêtre, phrase, paragraphe, document)
- Analyse distributionnelle (« à la Harris »)
 - regroupement de mots apparaissant dans les mêmes contextes syntaxiques
 - compléments des mêmes noms
 - adjectifs modificateurs des mêmes noms
 - syntagmes nominaux ou noms compléments des mêmes verbes
 - les classes ainsi construites doivent être validées, interprétées.
 - **nécessité d'une analyse syntaxique (robuste et partielle) préalable**

23/05/07

Gonisa Ninova IF09

53

IV. Un domaine d'application du TAL : Acquisition de RTO

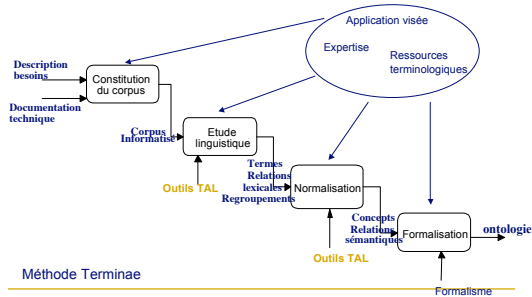


23/05/07

Gonisa Ninova IF09

54

IV. Un domaine d'application du TAL : Acquisition de RTO

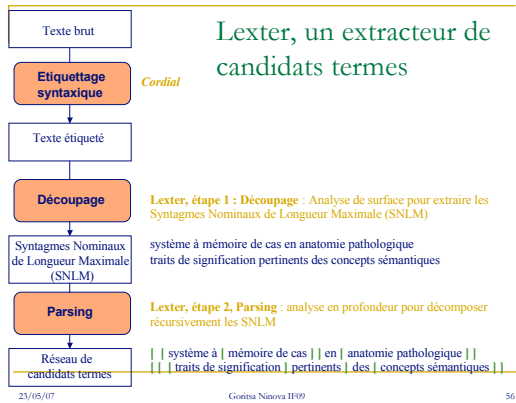


23/05/07

Gonisa Ninova IF09

55

Lexter, un extracteur de candidats termes



23/05/07

Gonisa Ninova IF09

56

La méthode Caméléon d'extraction de relations lexicales à partir de textes (Ségèla, 2001)

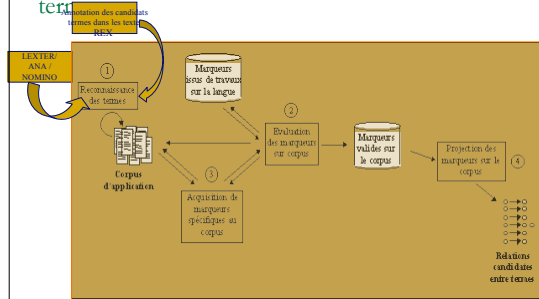
- **Objectif** : extraire des relations lexicales binaires entre termes à partir de texte technique
- **Base** : techniques de repérage de marqueurs linguistiques
- **Originalité** : permet la mesure et la prise en compte de la spécificité du corpus pour extraire un maximum de relation

23/05/07

Gonisa Ninova IF09

57

Méthode Caméléon d'extraction de relations entre termes



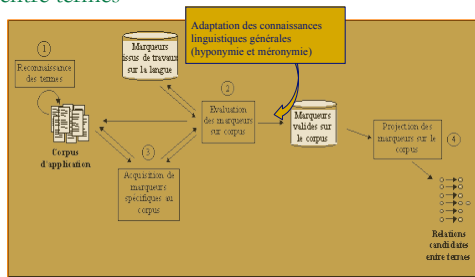
(Ségèla, 2001)

23/05/07

Gonisa Ninova IF09

58

La méthode Caméléon d'extraction de relations entre termes



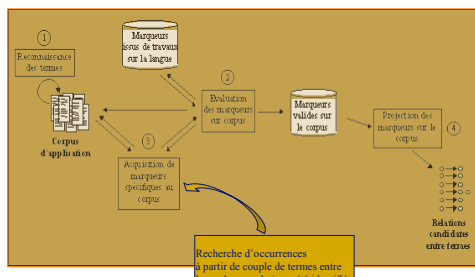
(Ségèla, 2001)

23/05/07

Gonisa Ninova IF09

59

La méthode Caméléon d'extraction de relations entre termes



(Ségèla, 2001)

23/05/07

Gonisa Ninova IF09

60

■ Emprunts à

- Nathalie Aussenac-Gilles (IRIT, Toulouse)
- Brigitte Biébow (LIPN, Paris)
- Didier Bourigault (ERSS, Toulouse)
- Patrick Séguéla (Synapse Développement, Toulouse)
- Sylvie Szulman (LIPN, Paris)

Bibliographie utilisée

- **Emile Benveniste**, 1966, *Problèmes de linguistique générale*, 1 III.10 Les niveaux de l'analyse linguistique
- **Pierrette Bouillon** (dir.), 1998, *Traitement automatique des langues naturelles*
- **Catherine Fuchs** (dir.), 1993, *Linguistique et Traitements Automatiques des Langues*
- **G. Ninova, A. Nazarenko, T. Hamon, S. Szulman**, 2005, *Comment mesurer la couverture d'une ressource terminologique pour un corpus ?*
- **Jean-Marie Prerrel** (dir.), 2000, *Ingénierie des langues*
- **Gilles Siouffi** (1999), *100 fiches pour comprendre la linguistique*
