Construction et interrogation de la structure informationnelle d'une base documentaire en français

Thèse de doctorat en Sciences du Langage - Linguistique et Informatique

Bernard Jacquemin

Université de la Sorbonne Nouvelle – Paris III
Xerox Research Centre Europe

8 décembre 2003

Introduction – Objectifs

- Des constatations
  - Une profusion de documents textuels ingérable
  - Un manque de structure dans l'information qu'ils contiennent
- Objectifs
  - identifier automatiquement l'information d'un énoncé
  - fournir un accès à cette information
  - structurer l'information textuelle pour la retrouver à la demande
  - utiliser pour ce faire des méthodes linguistiques
  - poser des requêtes en langage naturel pour en obtenir des réponses

Introduction – L'information

Question : *De quel chef Domitien est-il le successeur ?*

Document contenant le réponse : article *Domitien*:

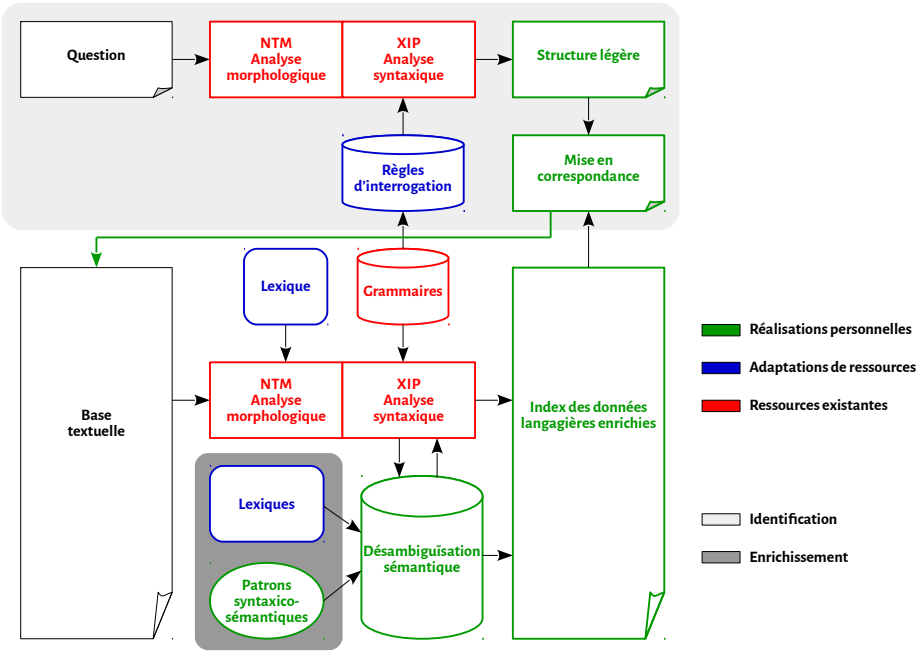
Second fils de Vespasien, il succéda à l'empereur Titus et poursuivit la remise en ordre de l'État.

- Une information lexico-syntaxique
  - nous ne sommes PAS dans une logique des prédicats sémantiques
  - l'information étudiée est constituée
    - des mots significatifs et de leur sens
    - des relations entre les mots qui mettent en rapport des éléments informatifs
    - des traits lexico-sémantiques impliquant le sens des mots ou de leurs relations
    - de toute paraphrase permettant de transformer la forme sans trop modifier le sens
  - l'information recherchée
    - mots et relations de la question
    - catégorisation de l'objet de la requête
    - granularité : la phrase

Plan

- Les enseignements des domaines de l'extraction d'information et de question-réponse
- Les acquis
  - la base textuelle : un banc d'essai réaliste
  - des outils d'analyse textuelle permettant de construire une structure
- Les adaptations et les apports personnels
  - la désambiguïsation sémantique lexicale
  - les ressources lexico-sémantiques
  - construction et enrichissement de la structure informationnelle
  - chercher et trouver une information en langage naturel
- Évaluation du système
- Conclusions et perspectives

Architecture du système

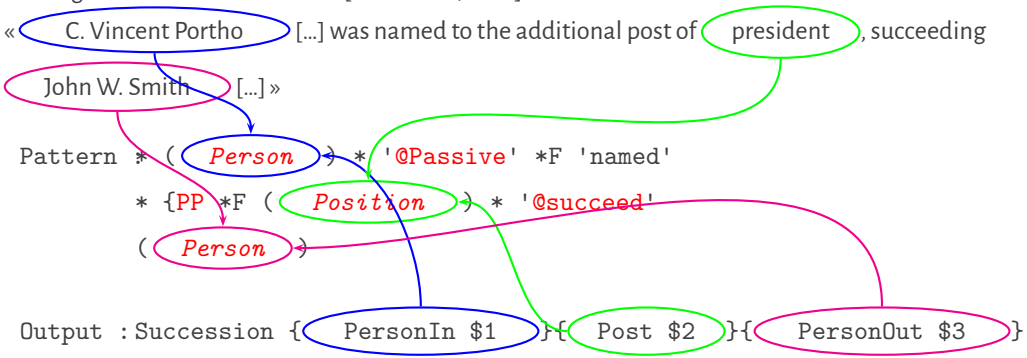


Enseignements de l'EI et de la tâche de QR (1/3)

- Méthode de mise en correspondance requête – réponse
  - EI (Extraction d'Information): ensemble des formes sous lesquelles apparaissent les fragments de l'information désirée dans les documents
  - QR (Question-Réponse): ensemble des formes sous lesquelles apparaissent les éléments de la question
- Les meilleurs systèmes introduisent des méthodes linguistiques complexes (LCCmain2002 = Falcon, limsiQalir2 = QALC dans TREC-11)
- Peu d'identification précise du sens liée au faible contexte de la requête
  $\Rightarrow$ Décision de traiter les documents à large contexte plutôt que les requêtes pour que l'expansion soit liée au sens de l'énoncé

Enseignements de l'EI et de la tâche de QR (2/3)

- Une règle d'extraction de WHISK [Soderland, 1999]

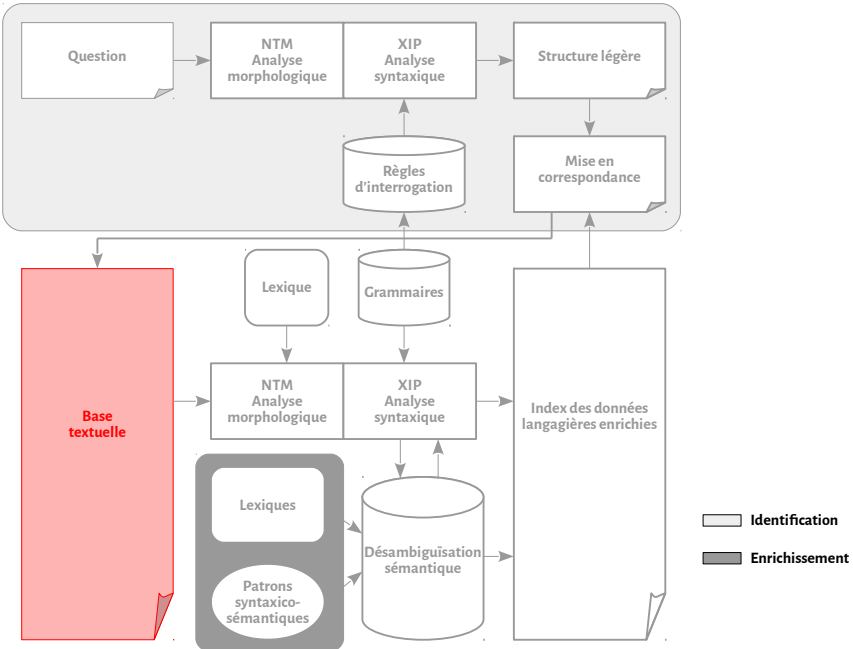


- AutoSlog [Riloff, 1996] ou CRYSTAL [Soderland et al., 1995] s'appuyaient déjà sur des éléments d'analyse linguistique

Enseignements de l'EI et de la tâche de QR (3/3)

- Question-réponse avec QUALC [Ferret et al., 2002 ; Monceaux, 2003]
  - catégorisation de la question (critères lexicaux, syntaxiques, sémantiques)
  - moteur de recherche avec utilisation de synonymes et de racinisation
  - extension des réponses candidates par des synonymes et des variations morphologiques
  - repérage des entités nommées (listes et règles)
    - expressions désignant des personnes, des organisations, des lieux…
    - expressions désignant des valeurs : dates, nombres, sommes…
- Depuis l'ébauche de [Hull, 1999] jusqu'au système de [Harabagiu et al., 2002], les techniques linguistiques sont de plus en plus prisées

La base textuelle utilisée



La base textuelle : un banc d'essai réaliste

- L'*Encyclopédie Hachette Multimédia*: un dictionnaire encyclopédique
- Des articles au format XML contenant du texte libre et semi-structuré
- Un balisage simple à exploiter et à enrichir
- Un texte soigneusement révisé (syntaxe, orthographe)
- Des sujets variés mais cohérents, qui se recoupent sans se recouvrir
- Un grand corpus
  - 75 000 articles dans autant de fichiers (40 000 encyclopédiques et 35 000 de langue)
  - taille des articles très variable (de quelques mots à plusieurs pages)
  - 90 Mo et 14,5 millions de mots, soit 194 mots par article en moyenne

La base textuelle : extrait

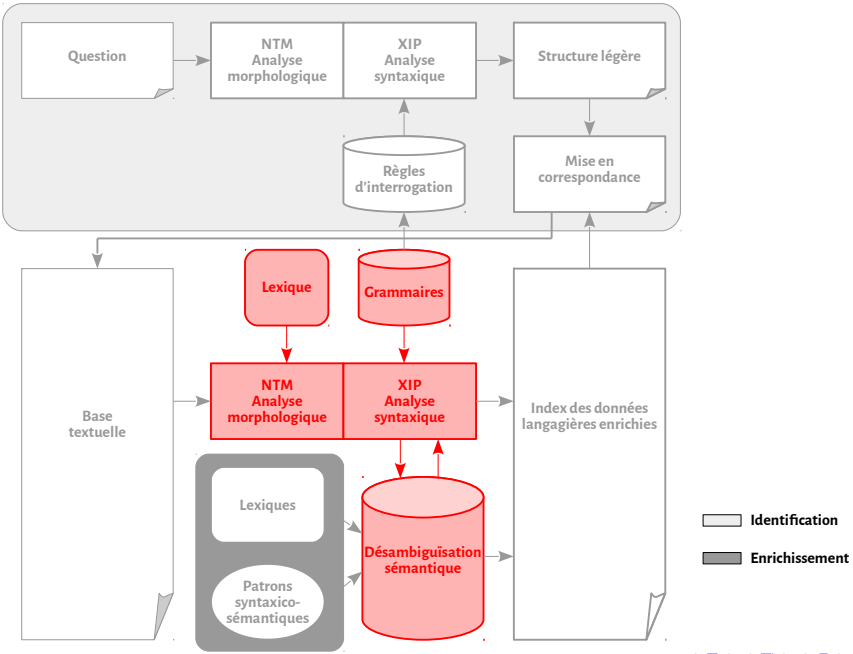
```
<Entree.ency>
    <Titre type="T">
        <Nom>Gunter</Nom>
        <Prenom>Edmund</Prenom>
    </Titre>
    <Resume>Astronome anglais</Resume>
    <Etatcv>(
        <Lieu type="naissance">dans le Hertfordshire</Lieu>
        <Date type="naissance">, 1581</Date> [...]
    </Etatcv>
```

En-tête

```
    <Parency>
        Il fut l'auteur de tables trigonométriques [...]
    </Parency>
</Entree.ency>
```

Document

Les outils d'analyse textuelle



Outils d'analyse textuelle (1/7)

- Construction d'une structure informationnelle
  - identification de l'information des documents
  - indexation de cette information documentaire
  - enrichissement de cette information pour la paraphraser au maximum sans pratiquement en modifier le sens
- Identification de l'information
  - identification des éléments d'information : les mots significatifs
  - identification des relations entre les éléments d'information : les relations syntaxiques entre mots significatifs
  - identification de la signification des éléments d'information : le sens des mots significatifs dans leur contexte

Analyse textuelle (2/7): morphologie

- Identification des mots : NTM (*Normalizer - Tokenizer - Morphological analyzer*) [Aït-Mokhtar, 1998]

Son deuxième fils [...]

Mot du texte	lemme	Analyse morphologique	Traits ajoutés
`son`	`son`	`+PP3S+InvGen+SG+Poss`	
`son`	`son`	`+Masc+SG+Noun+`	+SOM+AGR

`PP3S`	Pronom personnel 3e sg	`Masc`	Masculin
`InvGen`	Invariable en genre	`SG`	Singulier
`SG`	Singulier	`Noun`	Nom
`Poss`	Possessif	SOM	Relatif au corps
		AGR	Agriculture

- normalisation des mots (règles sous forme de transducteur)
- segmentation des mots (lexique sous forme de transducteur)
- analyse morphologique (transducteur à deux états)
- possibilités d'ajouts de données aux lexiques (étiquettes sémantiques)

Analyse textuelle (3/7): analyse syntaxique

- Exemple d'analyse syntaxique par XIP [Roux, 1999]

  Énoncé : « Il reconstruisit Rome ruinée par les incendies. »

  Extraction des dépendances :

`SUBJ(reconstruisit,Il)`	2e argument sujet du 1er argument
`SUBJ(ruinée,incendies)`	
`VMOD[INDIR](ruinée,par,incendies)`	3e argument compl. agent 1er argument
`VARG[DIR](reconstruisit,Rome)`	2e argument COD du 1er argument
`NMOD[ADJ](Rome,ruinée)`	2e argument épithète du 1er argument

  XIP effectue un découpage du texte en syntagmes minimaux, que nous n'exploitons pas

Analyse textuelle (4/7): analyse syntaxique

- Identification des relations : XIP (*Xerox Incremental Parser*)
  - ensemble de grammaires contextuelles
    - désambiguïsation catégorielle en fonction du contexte de chaque mot
    - construction des syntagmes minimaux : utilisation de traits sur les mots
    - construction des dépendances syntaxiques : utilisation de traits sur les mots, sur les syntagmes minimaux et sur les autres dépendances
  - ensemble de grammaires incrémentales
    - ajout aisé de traits sémantiques et autres
    - ajout aisé de dépendances syntaxiques ou non
    - exploitation des traits, notamment ceux fournis par l'analyse morphologique
    - utilisation d'une entité permettant d'assigner librement à un mot une information sous forme de trait

Analyse textuelle (5/7): sélection du sens

- Utilisation d'un désambiguïsateur sémantique pour
  - identifier le sens des mots polysémiques dans les documents
    - un système informatique voit des ambiguïtés là où un humain les résout (ambiguïté artéfactuelle) : ***il remporte [la victoire / les assiettes]***
    - importance du contexte syntaxique
    - utilisation efficace de dictionnaires et de corpus étiquetés
  - permettre un enrichissement focalisé sur le sens des mots

  $\Rightarrow$ besoin d'informations lexicales riches et cohérentes
- Extraction des règles de désambiguïsation à partir d'un dictionnaire
  - choix du sens selon le contexte syntaxico-sémantique de l'occurrence
    - schéma syntaxique (syntaxe uniquement) : ***il boit – il boit de l'eau***
    - schéma syntaxico-sémantique (syntaxe et sémantique) : ***embrasser quelqu'un*** $\neq$ ***embrasser quelque chose***
    - analyse syntaxique d'un exemple (lexique et syntaxe) : `VARG[DIR](remporter,victoire)`
    - généralisation de l'analyse d'un exemple (syntaxe et sémantique) : `VARG[DIR](remporter,[MIL])` (assaut, bataille)

Analyse textuelle (6/7): sélection du sens

Si ***remporter*** est la cible de la désambiguïsation et qu'il possède un objet direct qui porte le trait `MIL` (militaire), alors le sens choisi est « gagner »

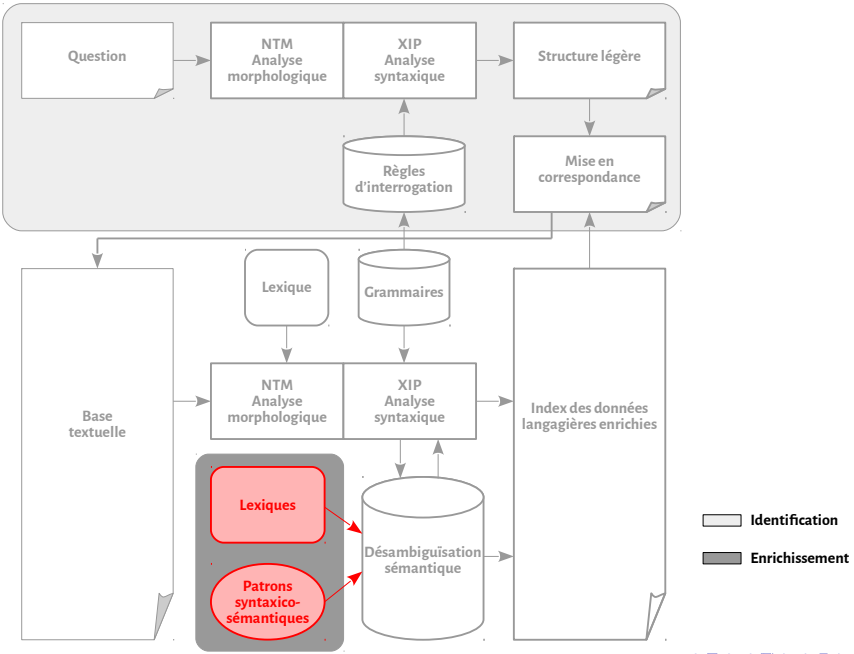
`remporter : VARG[DIR](remporter,[MIL])` $\Rightarrow$ `sens «gagner»`

- Construction d'une règle de désambiguïsation
  - le verbe « remporter » au sens « gagner » (***remporter une bataille***)
  - schéma syntaxique : verbe transitif direct
  - contrainte syntaxico-sémantique : objet direct de domaine MIL (militaire) issu de la généralisation de l'exemple ***On remporte la victoire sur ses adversaires***

Analyse textuelle (7/7): sélection du sens

- Application des règles de désambiguïsation
  - un ensemble de règles conditionnelles sous forme d'une grammaire XIP
  - exploitation du lexique identifié après la désambiguïsation catégorielle (remporter, victoire)
  - exploitation des dépendances de l'analyse syntaxique (`VARG[DIR](remporter,victoire)`)
  - exploitation des traits sémantiques attachés au mots (victoire `[MIL]`)
- Adjonction d'un trait de numéro de sens (correspondant à l'acception sélectionnée) pour les mots désambiguïsés (remporter `[NS2]` : gagner)
- Adjonction des traits sémantiques correspondant au sens choisi pour les mots désambiguïsés (remporter `[MIL]` : gagner)
- Une ambiguïté de sens peut subsister (remporter `[NS2]` un prix `[SPO]` ou remporter `[NS3]` un match `[SPO]`)

Les ressources lexico-sémantiques



Ressources lexico-sémantiques (1/4)

Des besoins liés à la désambiguïsation sémantique	**Dans *Dubois***
un lexique vaste et général	$\surd$
une information distribuée par acception	$\surd$
une information sémantico-syntaxique	$\surd$
des exemples pour chaque acception	$\sim \surd$

Des besoins liés à l'enrichissement	**Dans *Dubois***
des synonymes attachés au sens et non au mot	$\sim \surd$
des liens sémantiques entre catégories grammaticales (dérivation contrainte)	$\sim \surd$
une structure sémantique hiérarchique (hypéronymie, holonymie)	$\times$
des traits sémantiques (classes, domaines...)	$\surd$

$\Rightarrow$ Le dictionnaire ***Dubois et Dubois-Charlier*** répond à la plupart de ces exigences

Ressources lexico-sémantiques (2/4)

- Exemple : abaisser 01

Domaine	locatif, lieux
Classe	Transformation, changement - Sujet animé - Circonstanciel « avec »
Opérateur	rendre quelque chose bas ou devenir bas
Sens	baisser, descendre
Exemple	On abaisse le rideau de fer, le store. Le rideau s'abaisse vers le bas
Conjugaison	1^{e} conjugaison - paradigme : ***pleurer*** - auxiliaire avoir
Construction	Transitif direct - Sujet humain - COD inanimé - Circonstanciel instrumental
Dérivés	adjectifs en -ant (***abaissant***) - noms en -ment (***abaissement***) et en -eur (***abaisseur***)
Nom	dérivé préfixé à partir de ***baisse***
Lexique	dictionnaire de base 15 000 mots

Ressources lexico-sémantiques (3/4)

- Le dictionnaire *Dubois* manque de synonymes
  - récupération de listes dans *Bailly*, *Memodata*, *EuroWordNet*
  - un synonyme supplémentaire est ajouté à un sens du *Dubois* si les indices sémantiques du sens et du synonyme correspondent
  - exemple
    - le mot *ravir* au sens *voler*
    - synonymes fournis par les autres dictionnaires : *enlever* et *charmer*
    - *ravir*: domaine = SOC ; classe = S4
    - *enlever* au sens n° 6 (*priver de*) : domaine = SOC ; classe = S4
    - *charmer*: domaines PSY ou OCC ; classes P2 ou H2

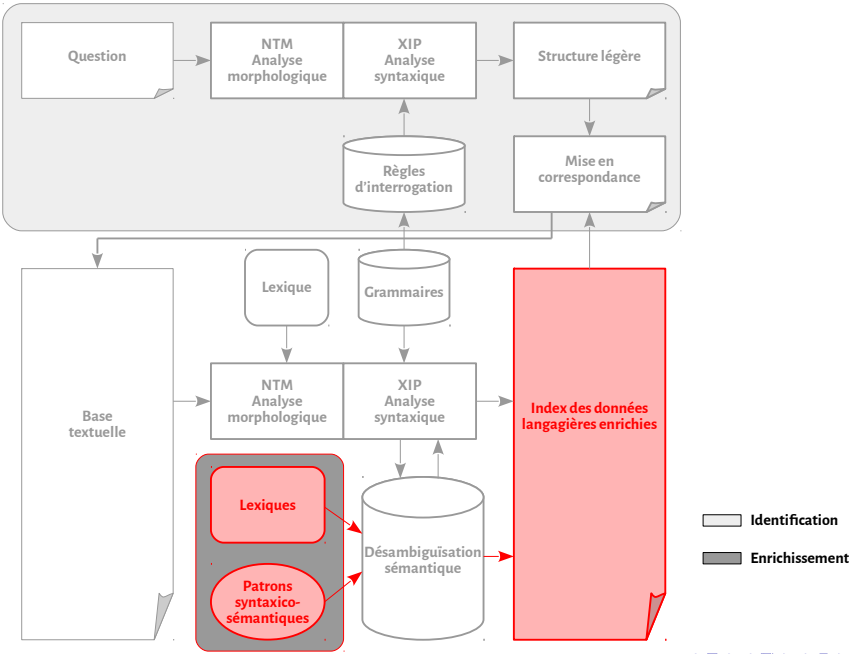
Ressources lexico-sémantiques (4/4): dérivation

- Les contraintes de dérivation relient un dérivé à un sens particulier du mot original
  - le dérivé *coupant* du verbe *couper*
    - correspond au sens *trancher* (n°1)
    - ne correspond pas au sens *interrompre l'électricité* (n°12)
- Les indications de dérivation dans le *Dubois* manquent de précision
  - génération de dérivés (existants) grâce à l'outil de [Gaussier, 1999]
  - élimination des inexistants grâce aux indications du *Dubois*
  - distribution par sens des dérivés grâce aux indications du *Dubois*

  Pour le verbe « couper »

Formes générées	Instruction *Dubois*	Numéro de sens *Dubois*
coup	–	suppression
coupure	dérivé nominal en *-ure*	1, 7, 9, 10, 12, 14, 16
coupable	–	suppression
coupant	adjectif verbal en *-ant*	1, 2
...	...	...

L'enrichissement de la structure

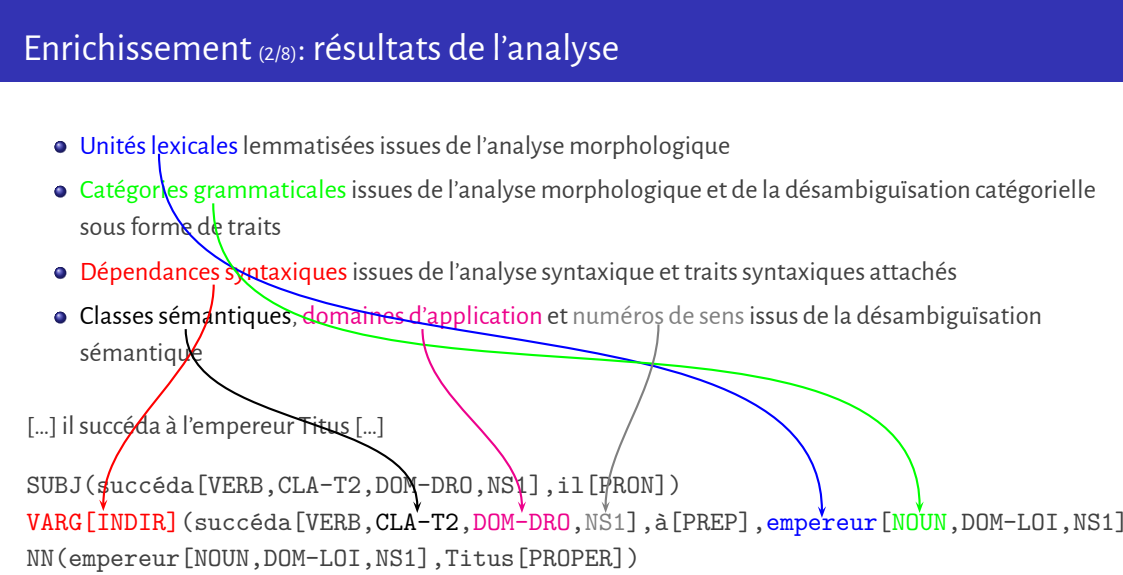


Enrichissement de la structure (1/8)

- Squelette de la structure informationnelle : les résultats des analyses
- Enrichissement synonymique par mot élémentaire
- Enrichissement synonymique par expression à mots multiples
- Enrichissement par dérivation morphologique contrainte
- Simulation d'un module de résolution de la coréférence des pronoms sujets

Enrichissement (2/8): résultats de l'analyse

- Unités lexicales lemmatisées issues de l'analyse morphologique
- Catégories grammaticales issues de l'analyse morphologique et de la désambiguïsation catégorielle sous forme de traits
- Dépendances syntaxiques issues de l'analyse syntaxique et traits syntaxiques attachés
- Classes sémantiques, domaines d'application et numéros de sens issus de la désambiguïsation sémantique



Enrichissement (3/8): synonymie (1/3)

Sélection des synonymes grâce au numéro de sens

- Synonymie simple : ajout des dépendances synonymiques dans la structure

Énoncé : « Son règne **favorise** la décadence » (synonyme : privilégier)

Dépendances originales	Dépendances synonymiques
`SUBJ(favorise, règne)`	`SUBJ(privilégier, règne)`
`VARG[DIR](favorise, décadence)`	`VARG[DIR](privilégier, décadence)`

Dépendances disjonctives

```
SUBJ(favorise OU privilégier, règne)
VARG[DIR](favorise OU privilégier, décadence)
```

Enrichissement (4/8): synonymie (2/3)

- Énoncé : « Il **commande** les **légions** »
- Synonymes
  - légion : troupe (simple)
  - commander : exercer son autorité sur (expression synonymique)
  - « Il **exercer son autorité sur** les légions »

Après synonymie simple

```
SUBJ(commande, il)
VARG[DIR](commande, légions OU troupe)
```

Après expression synonymique

```
SUBJ(commande OU exercer, il)
VARG[DIR](commande, légions OU troupe)
VARG[DIR](exercer, autorité)
NMOD[INDIR](exercer, sur, légions OU troupe)
```

Enrichissement (5/8): synonymie (3/3)

- Expressions synonymiques
  - remplacement du mot original par l'expression synonymique
  - analyse morpho-syntaxique de la phrase
  - construction de dépendances disjonctives (recouvrements)
  - élimination des dépendances redondantes

Enrichissement (6/8): dérivés

- Énoncé : « Pline **protégea** Suétone »
- Dérivé compatible : protecteur
- Schéma syntaxique original :
  VARG[DIR](*dérivable verbal*, X) (`VARG[DIR](protégea, Suétone)`)
- Schéma syntaxique correspondant :
  NMOD[INDIR](*dérivé nominal*, PREP, X) (`NMOD[INDIR](protecteur, PREP, Suétone)`)
- Correspondant sous forme textuelle :
  « Protecteur de Suétone »

Enrichissement (7/8): dérivés

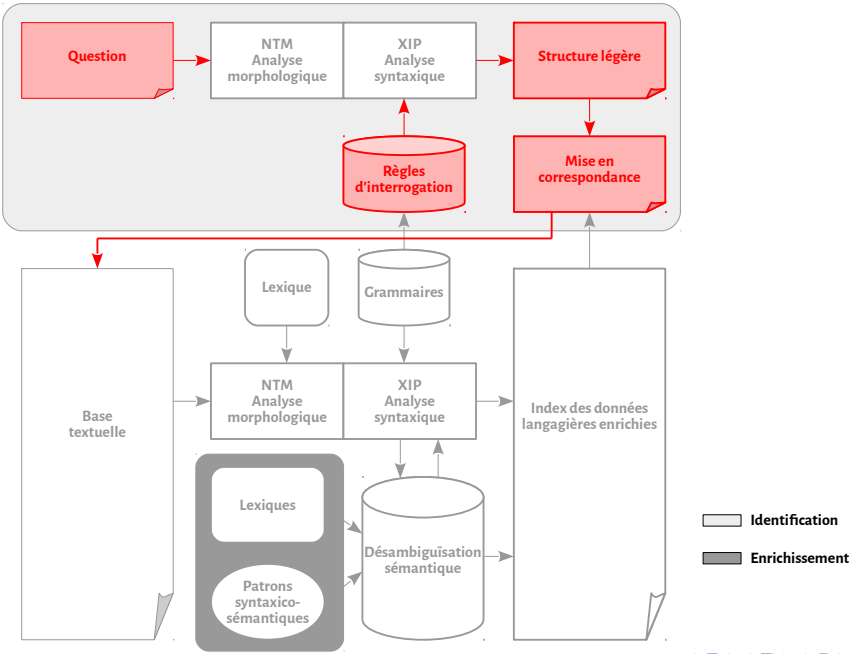
- Création d'une grammaire lexicale, syntaxique et sémantique à partir des patrons présents dans les tableaux de correspondance
- Application des règles fondée sur
  - le numéro de sens du mot considéré par l'enrichissement
  - le type de dérivation
  - la catégorie grammaticale du mot original et du dérivé
  - l'environnement syntaxique du mot considéré par l'enrichissement
- Génération de nouvelles dépendances où intervient le dérivé

  ou
- Insertion disjonctive du dérivé de même catégorie dans la structure

Enrichissement (8/8): simuler la coréférence

- L'intitulé d'un article encyclopédique est rarement repris dans le texte
  Domitien Second fils de Vespasien, il succéda à...
- Les articles sont structurés et le titre est aisément récupérable
  `<Entree.ency><Sommaire><Nom>Domitien</Nom></Sommaire>`
- Technique utilisée pour gérer le problème
  - tous les pronoms sujets de l'article font référence au titre
  - insertion disjonctive du titre dans les dépendances SUBJ impliquant un pronom
- exemple
  dépendance coréférente : `SUBJ(succède, il OU Domitien)`
- Cette simulation permet de résoudre l'absence de l'intitulé dans de nombreux articles

Trouver l'information



Trouver l'information (1/3)

- Construction d'une structure « légère » de la requête
  - analyse morpho-syntaxique de la requête (sans enrichissement)
  - catégorisation de l'objet de la question [Lehnert, 1979 ; Monceaux, 2003]
    - catégorisation de l'interrogatif
    - la dépendance syntaxique « `FOCUS` »
      - « `FOCUS` » est une relation de repérage de l'objet de la question (elle affecte le pronom interrogatif ou le nom qualifié par l'adjectif interrogatif)
      - Qui est le beau-père de Galère ? `FOCUS(beau-père[PAR])`
      - Qui combattit les Parthes ? `FOCUS(qui[humain])`
  - élimination des particularités syntaxiques liées à l'interrogation
    - suppression de l'interrogatif (maintien des traits sémantiques)
    - suppression des dépendances fonctionnelles
    - suppression de la dépendance syntaxique `FOCUS` (devient un trait)

Trouver l'information (2/3)

« De quel chef Domitien est-il le successeur ? »

~~SUBJ(est,Domitien)~~
~~FOCUS(chef)~~

NMOD[SPRED] (Domitien ,successeur)
NMOD[INDIR] (successeur ,de, chef [FOCUS,SOC,COM,HER,humain])

Document contenant la réponse : article *Domitien*
« […] il succéda à l'empereur Titus […] »

SUBJ(succéda,il)
VARG[INDIR](succéda,à,empereur)
NN(empereur,Titus)

SUBJ(succéda OU remplacer,il OU Domitien)
VARG[DIR](remplacer,empereur OU chef OU […])
NN(empereur OU chef OU souverain,Titus)

NMOD[INDIR] (successeur ,de, empereur OU chef OU […])
NMOD (il OU Domitien ,successeur)

Trouver l'information (3/3)

- Mise en correspondance de l'information présente dans la base textuelle et de celle demandée par la question
  - sélection de réponses candidates sur base d'une information plate
  - développement de la structure interrogative pour affiner les réponses
    - paramétrage sur le lexème ***focus***
    - paramétrage sur le nombre de dépendances en coïncidence (score)
    - utilisation ou non de la coréférence

Plan

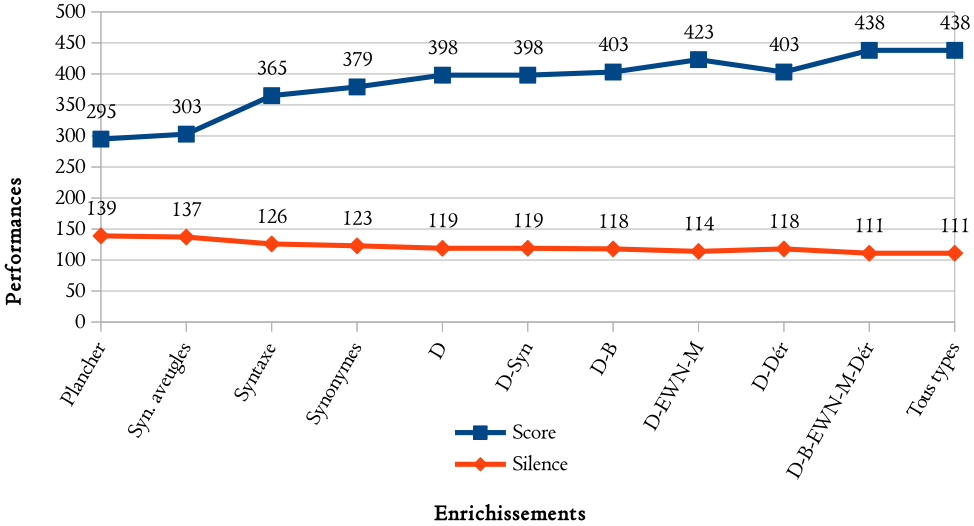
- Les enseignements des domaines de l'extraction d'information et de question-réponse
- Les acquis
  - la base textuelle : un banc d'essai réaliste
  - des outils d'analyse textuelle permettant de construire une structure
- Les adaptations et les apports personnels
  - la désambiguïsation sémantique lexicale
  - les ressources lexico-sémantiques
  - construction et enrichissement de la structure informationnelle
  - chercher et trouver une information en langage naturel
- Évaluation du système
- Conclusions et perspectives

Évaluation du système (1/6)

- Critères d'évaluation inspirés de la conférence TREC-8
  - Corpus de 50 articles, soit environ 20 000 mots
  - 200 questions attendant une réponse posées par 8 externes
  - QR : inverse du rang de la première bonne réponse : $score = \frac{1}{rang}$
  - EI : Précision, Rappel et F-mesure ($F = \frac{(\beta^2+1)PR}{\beta^2 P+R}$)
  - « Plancher » (*baseline*) : utilisation exclusive du lexique (sans enrichissement) de la structure informationnelle
- Différents paramétrages testés

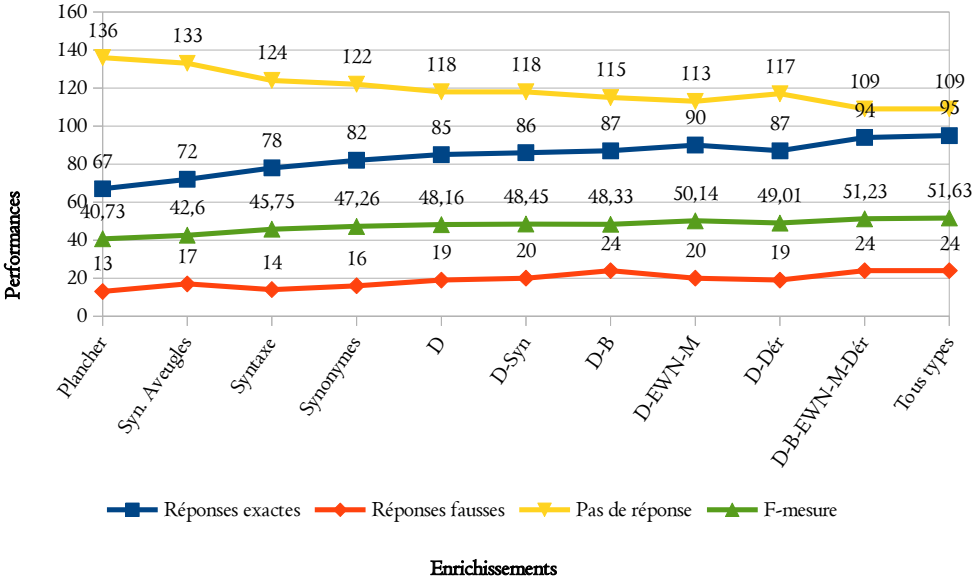
Évaluation du système (2/6)

- Efficacité de la méthode d'enrichissement



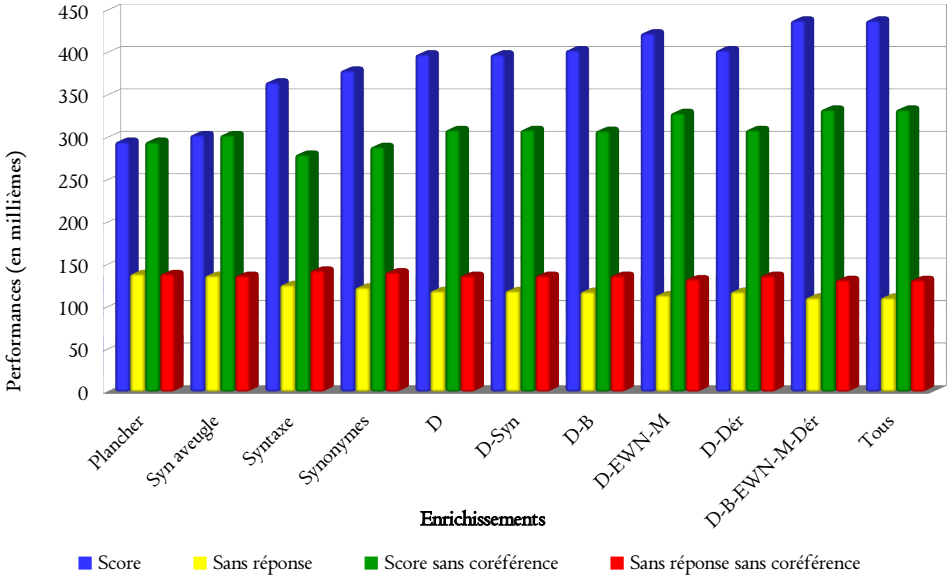
Évaluation du système (3/6)

- Faiblesse des résultats de la dérivation morphologique



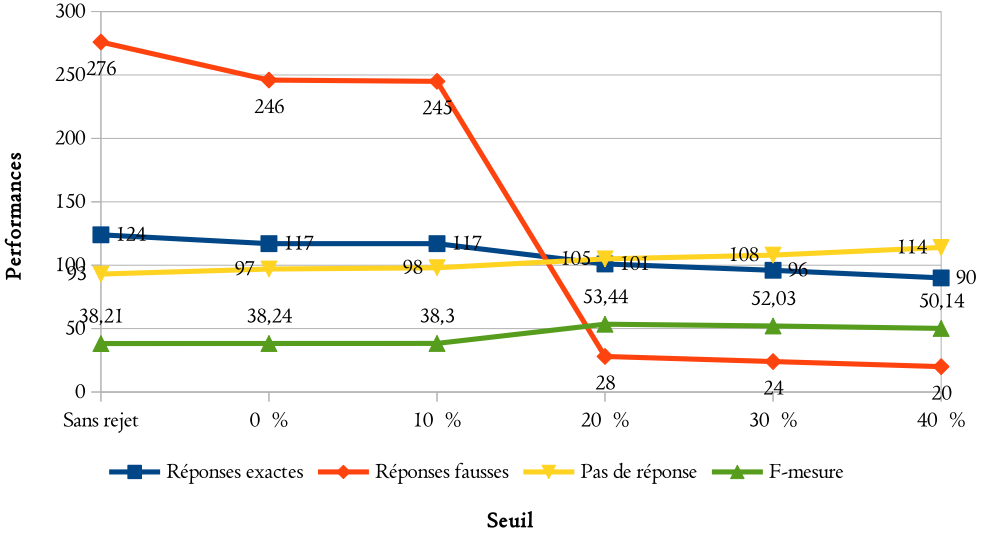
Évaluation du système (4/6)

- Grande importance de la coréférence



Évaluation du système (5/6)

- Aspect crucial du seuil de coïncidence en extraction d'information



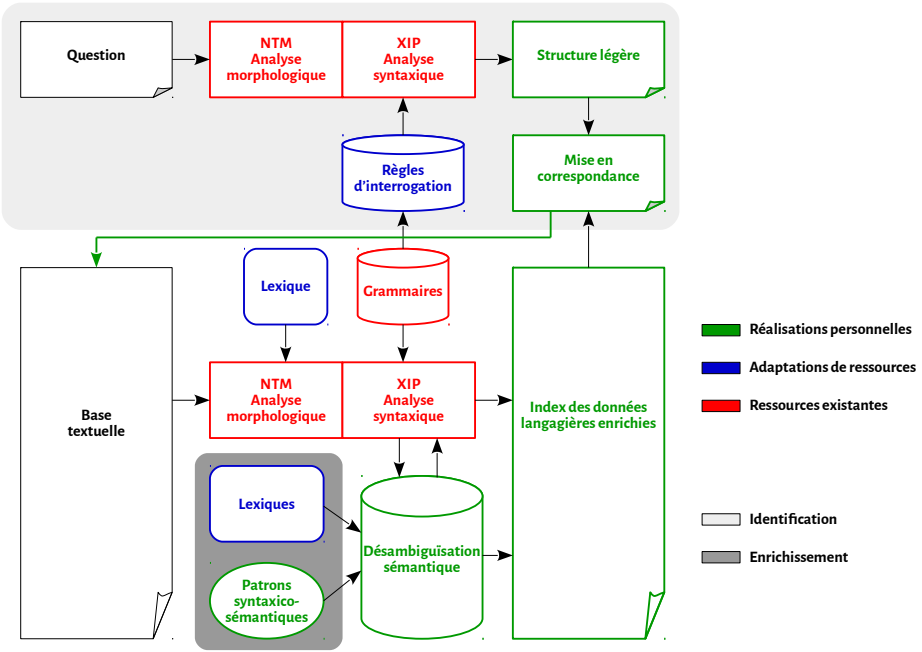
Évaluation du système (6/6)

- Synonymie encore trop pauvre
  - dictionnaires supplémentaires
  - généralisation nécessaire comme dans une taxinomie
  - constructions synonymiques (X fatal à Y = Y meurt de X)
- Semi-auxiliaires pris en compte dans le texte (se trouver, arriver...)
- Erreurs morphologiques et syntaxiques
- L'anaphore est un vrai problème
- Besoin d'un moteur logique et de meilleures connaissances du monde
- ...

Conclusion

- Analyse textuelle linguistique pour identifier l'information
  - identification des mots (analyse morphologique)
  - identification des relations (analyse syntaxique)
  - identification du sens des mots (désambiguïsation sémantique)
- Enrichissement de l'information lié au sens
  - synonymes et expressions synonymiques
  - dérivation morphologique et schémas syntaxique d'application
- Approche généraliste de l'information contenue dans des textes et de la manière de l'interroger

Architecture du système



Merci pour votre attention

Avez-vous des questions ?