Ressources linguistiques pour le TAL l'exemple des prédicats d'affect

Marne-la-Vallée 21 mai 2012

Pierre-André Buvet
Laboratoire LDI
UMR 7187 CNRS-UNIVERSITÉ PARIS 13
SORBONNE PARIS CITE

plan

1. Cadre théorique
   1.1. La théorie des 3 fonctions primaires
   1.2. La notion d'emploi prédicatif
   1.3. La linguistique de discours
2. Ressources linguistiques
   2.1 Les corpus
   2.2 Les dictionnaires électroniques
   2.3 Les grammaires locales
3. Applications
   3.1 Traitement automatique du discours rapporté
   3.2 Résolution automatique des anaphores
   3.3 Catégorisation des textes en fonction de leur degré de subjectivité

Cadre théorique

1) observer les faits de langue étudiés pour en dégager des régularités généralisables

2) exprimer ces régularités en éliminant le flou, l'implicite, le non-dit, les évidences allant de soi

3) vérifier la cohérence de la formulation qui garantit l'objectivité et donc la reproductibilité de la démarche

La théorie des trois fonctions primaires postule que les énoncés minimaux sont des structures prédicat-argument actualisées.

Ce garçon détestait son père

prédicat	*détest-*
arguments	*garçon* et *père*
actualisateurs	*-ait*, *ce*, *son*

Cadre théorique

Il est posé que les prédicats prédominent les arguments car les occurrences des arguments sont subordonnées à celles des prédicats

Par contre, les occurrences des prédicats ne procèdent pas nécessairement de celles de leurs arguments

Cadre théorique

La structure prédicat-argument conduit à distinguer les actualisateurs selon qu'ils se rapportent aux prédicats ou aux arguments

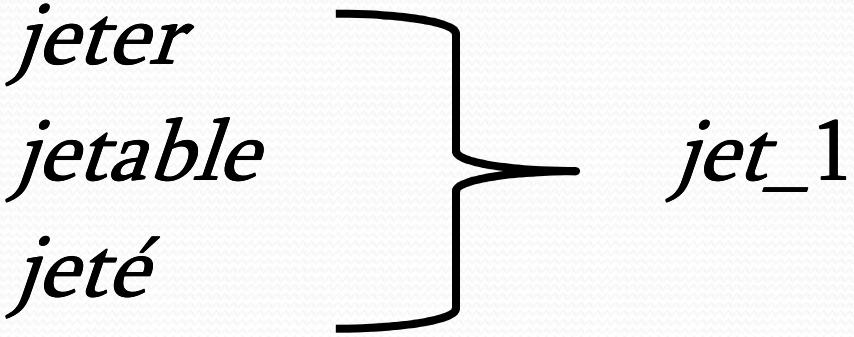
Dans le premier cas, les occurrences des actualisateurs dépendent uniquement de celles des prédicats, par ex. l'indication temporelle est imputable au seul verbe.

Dans le second cas, elles dépendent des relations entre les prédicats et leurs arguments, la combinatoire des noms avec le verbe est un des éléments d'explication de leur combinatoire avec le déterminant.

La catégorisation syntactico-sémantique ne correspond pas à la catégorisation grammaticale. Par exemple, les noms s'interprètent comme :

- des prédicats (*faim* dans *Luc a faim*) ;
- des arguments élémentaires (*olive* dans *Luc mange une olive*) ;
- des actualisateurs (*kyrielle* dans *Luc a lu une kyrielle de romans*).

Un prédicat est caractérisé par une racine prédicative, une classe sémantique et un domaine d'arguments



Un emploi prédicatif est caractérisé *a minima* par :

- une racine prédicative
- une classe sémantique
- un type sémantique
- un aspect inhérent
- une construction
- une distribution morphosyntaxique
- une distribution sémantique

jeter

- jet_1 (racine)
- DEBARRAS (classe sémantique)
- action (type sémantique)
- duratif perfectif (aspect inhérent)
- X0 V X1 (PREP2) X2 (construction)
- X0 = GN/X1 = GN /X2 = GN (distribution 1)
- X0 =HUMAIN/X1 = OBJET/X2 = LIEU (distribution2)

jeter

- jet_1 (racine)
- DEBARRAS (classe sémantique)
- action (type sémantique)
- duratif imperfectif (aspect inhérent)
- X0 V (construction)
- X0 = GN (distribution 1)
- X0 =HUMAIN (distribution2)

jetable

- jet_1 (racine)
- DEBARRAS (classe sémantique)
- état (type sémantique)
- permanent (aspect inhérent)
- X0 être A (construction)
- X0 = GN (distribution 1)
- X0 =OBJET (distribution2)

jeté

- jet_1 (racine)
- DEBARRAS (classe sémantique)
- état (type sémantique)
- résultant (aspect inhérent)
- X0 être A (construction)
- X0 = GN (distribution 1)
- X0 =OBJET (distribution2)

Cadre théorique

La linguistique du discours se caractérise par une approche intégrée telle que les différentes facettes de l'analyse sémantique sont considérées comme se rapportant conjointement au lexique, à l'énonciation et à la compréhension.

Il ne s'agit pas d'une accumulation de traitements sémantiques composites mais d'un traitement homogène qui prend appui sur le lexique pour analyser les autres phénomènes langagiers.

Trois niveaux discursifs :

- le niveau logico-sémantique
- le niveau énonciatif
- le niveau interprétatif

Quelques chiffres à propos des prédicats d'affect :

≈ 400 racines prédicatives

≈ 900 emplois verbaux

≈ 500 emplois nominaux

≈ 700 emplois adjectivaux

≈ 200 emplois verbes à caractère causatif

≈ 300 adjectifs à caractère causatif

Cadre théorique

ADMIRATION	*émerveill-*
AMOUR	*attach-*
COLERE	*courrou-*
CONTRARIETE	*ennuy-*
ENTHOUSIASME	*emball-*
HAINE	*détest-*,
JOIE	*euphori-*,
MEPRIS	*déconsidér-*,
PEUR	*paniqu-*

…

Classe sémantique PEUR

baliser, mouiller, horrifié, effarement, effarer, effaré,effarouchement, effaroucher, effroi, effrayer, s'effrayer, effrayé,épouvante,épouvanter, épouvanté, frayeur, frousse, panique, paniquer, paniqué, pétoche, pétocher, peur, apeurer, apeuré, terreur, terrifier, terrifié, terreur, terroriser, terrorisé, trac, trouille

Classe sémantique COLERE

fumasse, se déchaîner, déchaîné, se fâcher, fâché, colère, courroux, courroucer, courroucé, emportement, s'emporter, fureur, furieux, furibard, furibond, furax, rage, enrager, enragé, rage, rager, rogne

Cadre théorique

« La notion de modalité implique l'idée qu'une analyse sémantique permet de distinguer, dans un énoncé, un *dit* (appelé parfois 'contenu propositionnel') et une *modalité* - un point de vue du sujet parlant sur ce contenu »

Jean Cervoni *L'énonciation* PUF

triste : emploi prédicatif de la sous-catégorie AFFECT_NEGATIF de la catégorie DESCRIPTION SUBJECTIVE

On pleure quand on est triste (assertion)

C'est triste de faire cela (modalité élocutive)

applaudir : emploi prédicatif de la sous-catégorie ECHANGE de la catégorie DESCRIPTION INTERINDIVIDUELLE

La presse a applaudi sa performance (assertion)

Je vous applaudis des deux mains (modalité allocutive)

Un petit tour sur le site officiel du Festival plus tard, je suis immédiatement soulagé : il y a bien un endroit pour déposer ses valises. Ouf ! Me voilà donc parti pour un court trajet en train entre Antibes et Cannes. Ensuite, dépôt de valise, retrait de mon accréditation presse (la jaune, la « moins » prestigieuse mais après tout, peu importe, l'important étant d'être là), passage au bureau de presse pour retirer mes invitations pour le Théâtre Lumière : 15 minutes montre en main. Je suis donc finalement très large pour le Sleeping Beauty de l'australienne Julia Leigh. L'émotion n'est évidemment plus la même cette année au moment de gravir les fameuses marches et de m'installer dans la grande salle, c'est plutôt de la joie de me retrouver à nouveau plongé au cœur de ce tourbillon permanent.

Ressources linguistiques

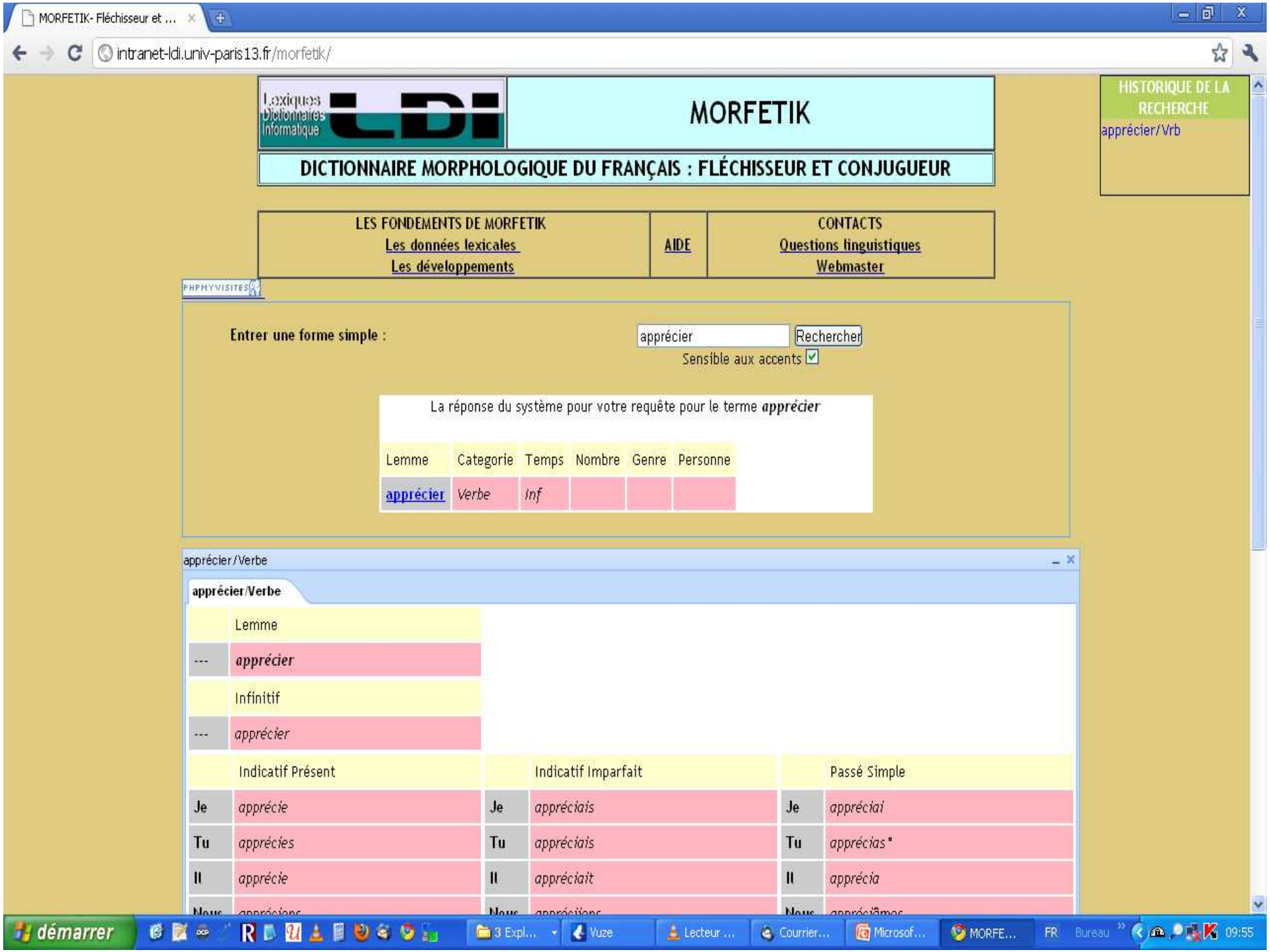
- les méthodes statistiques sont souvent privilégiées au dépend des méthodes linguistiques en TAL

- néanmoins, les traitements purement statistiques semblent avoir atteint un point critique

- il s'ensuit que, depuis quelques années, des méthodes mixtes sont utilisées

Les dictionnaires électroniques:

- dictionnaire morphosyntaxique :
  - MORFETIK simple
  - MORFETIK complexe
- dictionnaires syntactico-sémantiques
  - PRED-DIC
  - ARGU-DIC
  - ACTU-DIC
  - ETHU_DIC


MORFETIK- Fléchisseur et ...
intranet-ldi.univ-paris13.fr/morfetik/
Lexiques Dictionnaires Informatique
LDI
MORFETIK
DICTIONNAIRE MORPHOLOGIQUE DU FRANÇAIS : FLÉCHISSEUR ET CONJUGUEUR
HISTORIQUE DE LA RECHERCHE
apprécier/Vrb
LES FONDEMENTS DE MORFETIK
Les données lexicales
Les développements
AIDE
CONTACTS
Questions linguistiques
Webmaster
PHPMYVISITES
Entrer une forme simple :
apprécier
Rechercher
Sensible aux accents
La réponse du système pour votre requête pour le terme apprécier
Lemme
Categorie
Temps
Nombre
Genre
Personne
apprécier
Verbe
Inf
apprécier/Verbe
apprécier/Verbe
Lemme
---
apprécier
Infinitif
---
apprécier
Indicatif Présent
Indicatif Imparfait
Passé Simple
Je
apprécie
Je
appréciais
Je
appréciai
Tu
apprécies
Tu
appréciais
Tu
apprécias
Il
apprécie
Il
appréciait
Il
apprécia
démarrer
3 Expl...
Vuze
Lecteur ...
Courrier...
Microsof...
MORFE...
FR
Bureau
09:55

Ressources linguistiques

```
use warnings;
use Encode;
$fic = $ARGV[0];
open (IN, '<:encoding(utf-8)', $fic) or die;

# Suppression des mots à exclure à partir du fichier "dlf_exclusion_list.txt" :

open (INE, '<:encoding(utf-8)', "dlf_exclusion_list.txt");
open (INA, '>:encoding(utf-8)', $fic.".clean");

while (<INE>){
        push (@tab, $_);
        }

while (<IN>){
        $line = $_;
        next if ($line =~ /PADV/);
        next if ($line =~ /Profession/);
        next if ($line =~ /Advconjs/);
        next if ($line =~ /Prépconjs/);
        next if ($line =~ /VADV/);
        next if ($line =~ /PFX/);
        next if ($line =~ /PCPN/);
        next if ($line =~ /XI/);
        $drapeau = 0;
        foreach $tab (@tab){
                if ($tab eq $line)
                        {
                $drapeau = 1;
                        }
                }
        if ($drapeau != 1) {
                print INA $_;
#               print INA $_."\r\n";
                }
        }
```

Ressources linguistiques

- a,.N:ms:mp
- a,.N+z1:ms:mp
- alpes,alper.V:P2s:S2s
- autour,.N:ms
- b,.N+z1:ms:mp
- badge,badger.V+z2:P1s:P3s:S1s:S3s:Y2s
- bénéficier,.N+z1:ms
- bien,.ADV+PADV+z1
- bien,.ADV+VADV+z1
- boutique,boutiquer.V:P1s:P3s:S1s:S3s:Y2s
- c,.N+z1:ms:mp
- ç,.N+z1:ms:mp
- ça,.N+Abst+z1:ms
- ça,.PRO+Ton+z1:3ms
- carte,carter.V+z2:P1s:P3s:S1s:S3s:Y2s
- ce,.PRO+z1:3s
- certains,certain.N:mp
- d,.N+z1:ms:mp
- dire,.N+z1:ms
- disques,disquer.V:P2s:S2s
- doit,.N:ms
- donc,.ADV
- dossiers,dossier.N+z1:mp
- e,.N+z1:ms:mp
- enceintes,enceindre.V+z2:Kfp
- enceintes,enceinter.V:P2s:S2s
- est,.N+z1:ms
- est,.A+z1:ms:fs:mp:fp
- étaient,étayer.V+z1:P3p:S3p
- évident,évider.V+z2:P3p:S3p
- f,.N+z1:ms:mp
- face,facer.V:P1s:P3s:S1s:S3s:Y2s
- faites,faire.V+z1:P2p:Y2p
- force,.DET+'Dadj'+z1:ms:fs:mp:fp
- foyer,.A+z2:ms
- g,.N+z1:ms:mp
- grues,gruer.V:P2s:S2s
- h,.N+z1:ms:mp
- housse,housser.V:P1s:P3s:S1s:S3s:Y2s
- i,.N+z1:ms:mp
- impact,.A+z1:ms:fs:mp:fp
- j,.N:mp:ms
- j,.N+z1:ms:mp
- jours,.N:mp
- k,.N+z1:ms:mp
- l,.N+z1:ms:mp
- la,.N+[Mus]+z1:ms:mp
- la,.N+z1:ms:mp
- là,.INTJ+z1
- léonard,.A:ms
- lettre,lettrer.V:P1s:P3s:S1s:S3s:Y2s
- leurs,.N+HumColl+z1:mp
- lié,.N:ms
- lignes,ligner.V:P2s:S2s

Ressources linguistiques

breloque	,.N	H_VETEMENT	C_VETEMENT_PARURE
breloques	,.N	H_VETEMENT	C_VETEMENT_PARURE
camisole	,.N	H_VETEMENT	C_VETEMENT_HABIT
camisoles	,.N	H_VETEMENT	C_VETEMENT_HABIT
centrale vapeur	,.N	H_APPAREIL	H_APPAREIL_REPASSAGE
centrales vapeur	,.N	H_APPAREIL	H_APPAREIL_REPASSAGE
centrales vapeurs	,.N	H_APPAREIL	H_APPAREIL_REPASSAGE
chaussée	,.N	H_LIEU	C_VOIE
chevalière	,.N	H_VETEMENT	C_VETEMENT_PARURE
chevalières	,.N	H_VETEMENT	C_VETEMENT_PARURE
corne	,.N	H_VETEMENT	C_VETEMENT_PARURE
cornes	,.N	H_VETEMENT	C_VETEMENT_PARURE
croix	,.N	H_VETEMENT	C_VETEMENT_PARURE
faux=palier	,.N	H_DISPOSITIF	H_DISPOSITIF_RANGEMENT
garde=corps	,.N	H_ORGANE	H_DISPOSITIF_PROTECTION

ɔi - iété	*couper*-1	*couper*-2	*coupure*-1	*coupure*-2
	coup-	*coup*-	*coup*-	*coup*-
	SEPARATION	SEPARATION/ ENTAILLE	ENTAILLE	ENTAILLE
	action	état	action	état
	duratif accompli	permanent	duratif accompli	provisoire
	X0 V X1 (PREP2 X2)	X0 V	X0 faire DET N PREP1 X1 (PREP2 X2)	X0 avoir DET N PREP1 X1
	X0= GN X1 = GN	X0 = GN X0 = GN	X0 = GN X1 = GN X2 = GN	X0 = GN X1 = GN
	X0 = HUMAIN X1 = $OBJET_1$ X2 = $OBJET_2$	X0 = $OBJET_2$	X0 = HUMAIN/OBJET2 X1 = HUMAIN X2= PARTIE_CORPS	X0 = HUMAIN X1= PARTIE_CORPS

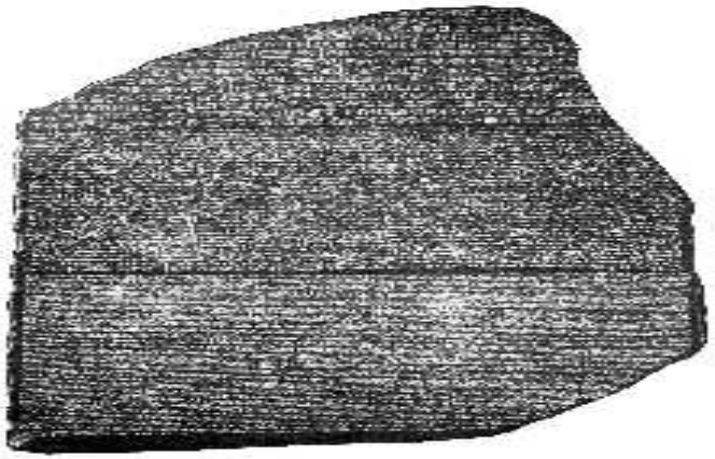
Interface d'Interrogation des Prédicats Adjectivaux

Lexiques
Dictionnaires
Informatique

IIPA

ENTREZ

conception : Pierre-André Buvet LDI
ressources : Pierre-André Buvet LDI



Ressources linguistiques

Les corpus exploités en Traitement Automatique des Langues sont généralement des textes bruts ou des textes bruts annotés.

On peut les distinguer selon leur finalité : corpus de travail, corpus d'apprentissage, corpus d'expérimentation, corpus d'évaluation, etc.

L'élaboration d'outils performants pour effectuer les analyses linguistiques est fondée sur l'exploitation de trois sortes de corpus de travail:

(i) le corpus d'investigation ;

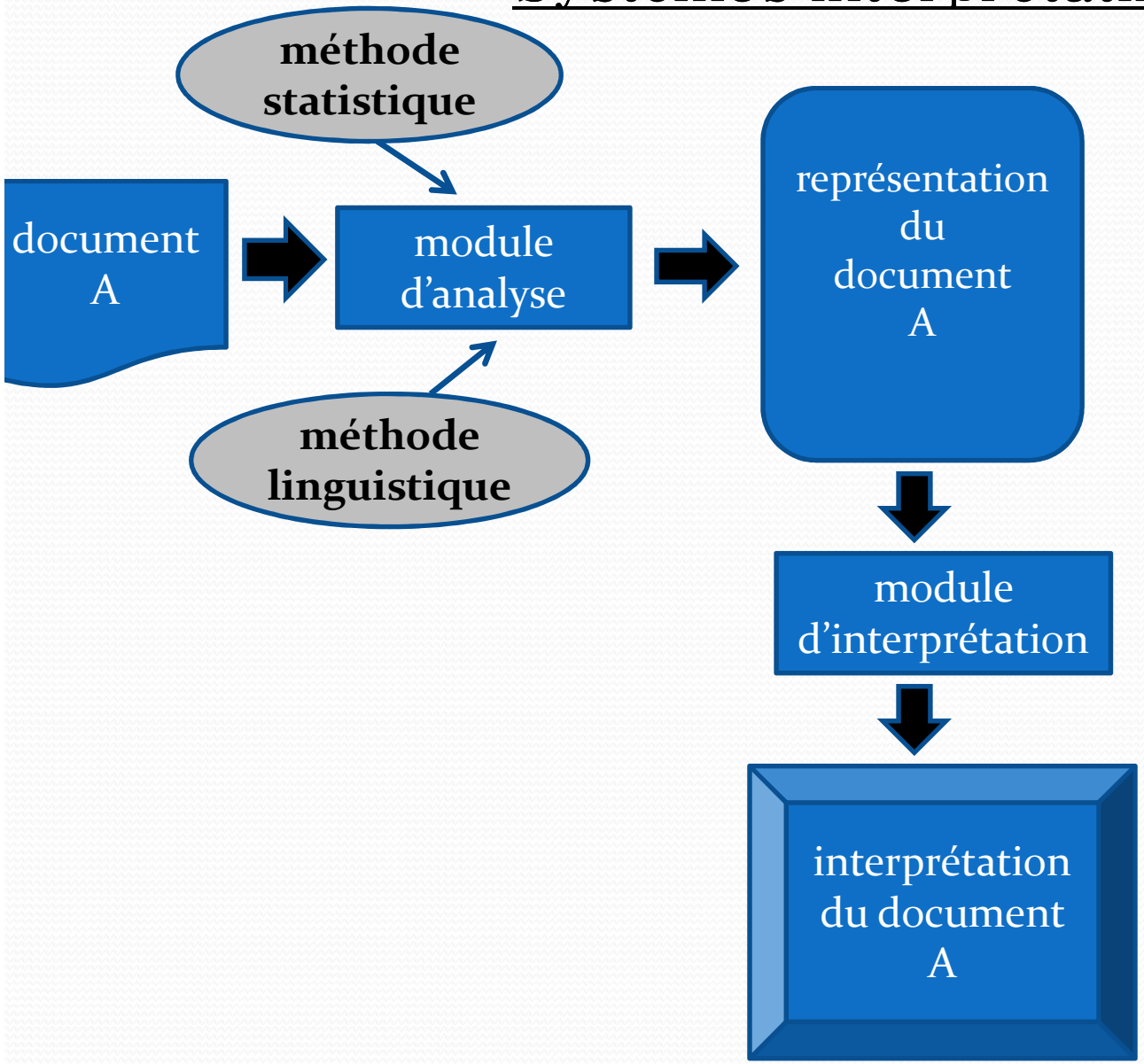
(ii) (ii) le corpus de test ;

(iii) (iii) le corpus de validation.

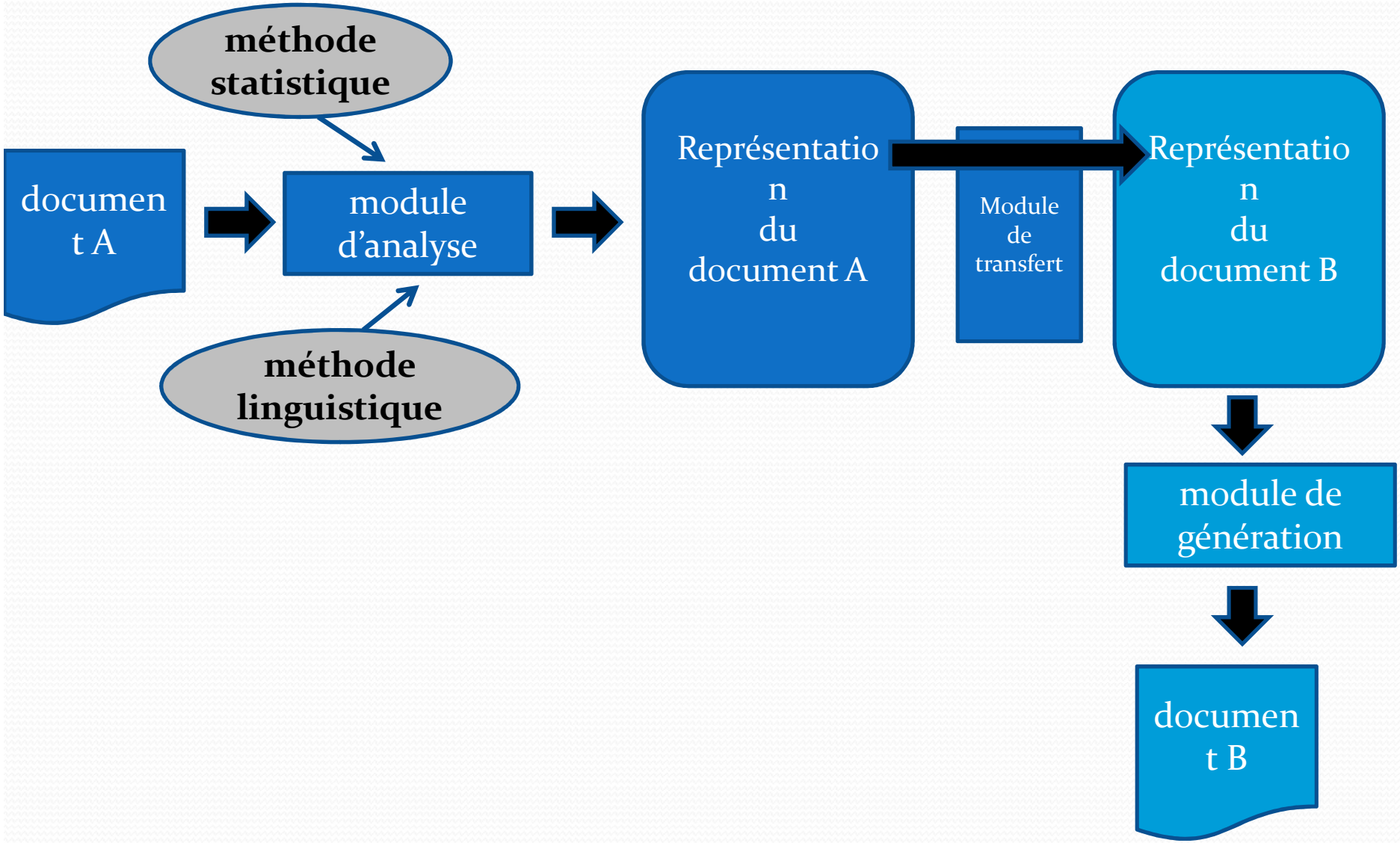
Applications

- les applications sont toutes fondées sur des représentations des documents
- les représentations peuvent résulter d'une phase d'analyse
- les représentations peuvent aussi être fournies en entrée à des systèmes qui comportent uniquement une phase de génération de documents

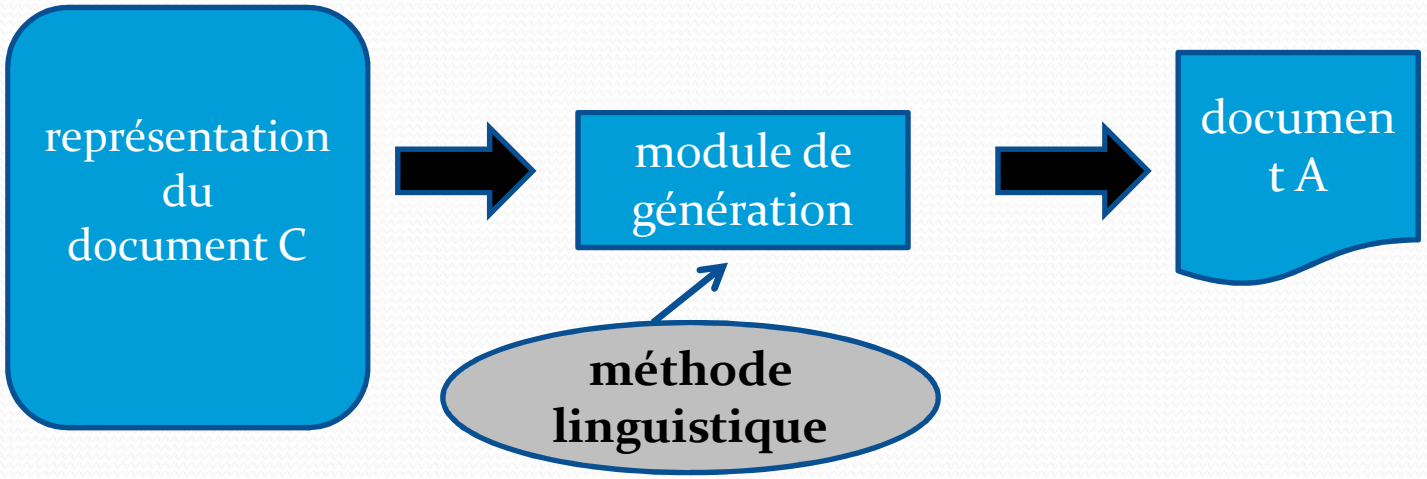
Systèmes interprétatifs



Systèmes interprétatifs-transformationnels



Systèmes créatifs



Applications

Systèmes interprétatifs
Moteur de recherche
Système de Veille
Système d'aide à la décision
Système d'aide à la communication
Plate-forme collaborative

Applications

Systèmes interprétatifs-transformationnels
Traduction Automatique
Résumé Automatique
Correction Orthographique
Correction Grammaticale

Applications

Systèmes créatifs
Compte rendu automatique
Producteur de lettres personnalisées

Applications

- TRAITEMENT AUTOMATIQUE DU DISCOURS RAPPORTE
- RÉSOLUTION AUTOMATIQUE D'ANAPHORE
- DETECTION AUTOMATIQUE DE LA SUBJECTIVITE

Recherche sur le site

LDI (Lexiques, Dictionnaires, Informatique) UMR 7187

Le **LDI** : « Lexiques, Dictionnaires, Informatique » est un laboratoire qui part du lexique pour élaborer ou analyser des dictionnaires en utilisant l'informatique.

- Le **lexique** représente le matériau de base qui sert d'entrée à la description des langues.

- Le **dictionnaire** est à la fois une source de données et un objet construit.

- L'**informatique** est un outil mis à la disposition des linguistes pour analyser ou élaborer les dictionnaires et pour concevoir des méthodologies appropriées au traitement automatique des langues.*(Lire la suite)*

Nos ressources

- Musée virtuel des dictionnaires
- Noms composés
- Dictionnaire du Français Scientifique Médiéval
- Corpus Droits de l'Homme
- Morfetik
- le Petit Larousse Illustré de 1905
- dBox

Laboratoire
- Page d'accueil
- Membres du LDI
- Doctorants
- Thèses soutenues
- Séminaires
- Contacts
- Nous rejoindre

Nos publications
- Nos publications
- Rapports d'activité
- Ouvrages

Formations
- Master Pro TILDE
- Master Recherche SCIL

Revues
- Neologica
- Cahiers de lexicologie

Coopérations
- Les partenaires

Séminaires LDI

Signature du co-diplôme de formation MasterproTilde 2 entre l'Université de Cracovie , Pologne et l'Université de Paris 13.

27 mai 2011
Journée scientifique conscila
La dénomination : perspectives linguistique

15 et 16 septembre 2011

Journées scientifiques LTT, Villetaneuse (Paris 13)

Appel à candidature pour un sujet de thèse pour septembre 2011

Sujet :

Modalité et inférence

Résolution des anaphores nominales pour la compréhension automatique des textes

pabuvet@ldi.univ-paris13.fr