Exploiter la structure discursive pour la recherche automatique de segments obsolescents

Marion Laignelet

IRIT

08 décembre 2009

Problématique générale

Une situation concrète

existence de documents volumineux et riches en informations : les encyclopédies

Un besoin réel/industriel

mettre à jour ce type de documents

$\Rightarrow$ peut-on décider de façon automatique si une information doit être mise à jour ?

Un exemple de segment à mettre à jour

x. Actualité

§ Établir une liste exhaustive des avancées récentes de la recherche médicale est impossible tant les progrès sont nombreux. Toutefois, il convient de rappeler un certain nombre de découvertes très récentes. En 2003, l'une des grandes priorités de la recherche médicale internationale a concerné le sida.

Source : Corpus ATLAS (fiche Médecine - Le Sida)

Date de publication : 2003

Un exemple de segment à mettre à jour

Exploiter des indices linguistiques ?

x. **Actualité**

§ Établir une liste exhaustive des **avancées récentes** de la recherche médicale est impossible tant les **progrès** sont nombreux. Toutefois, il convient de rappeler un certain nombre de **découvertes très récentes**. **En 2003,** l'une des grandes priorités de la recherche médicale internationale a concerné le sida.

Source : Corpus ATLAS (fiche Médecine - Le Sida)
Date de publication : 2003

Définition de la tâche de mise à jour

notion d'*obsolescence*

repérer les zones textuelles dans lesquelles l'information est susceptible d'évoluer

Un double objectif

- décrire linguistiquement les segments d'obsolescence
- proposer un prototype d'aide à la mise à jour

Une méthodologie ancrée dans :

- une **linguistique discursive**
- une **linguistique de corpus**
- une **linguistique outillée**

$\Rightarrow$ hypothèse forte = les indices de type structurels (titres, adverbiaux détachés et positions) sont centraux

Plan

Une méthode en corpus

Plan

Plan

Une méthode en corpus

Linguistique et discours

Repérage et annotation automatiques des indices potentiels de l'obsolescence

Plan

Une méthode en corpus

Linguistique et discours

Repérage et annotation automatiques des indices potentiels de l'obsolescence

Description et analyse statistique

Plan

Une méthode en corpus

Linguistique et discours

Repérage et annotation automatiques des indices potentiels de l'obsolescence

Description et analyse statistique

Conclusion et Perspectives

Plan

Des textes issus du monde professionnel : le corpus [ENCYCLO]

Taille : 10 000 phrases

Constitution : divisé en deux sous-corpus :

- sous-corpus [ATLAS]
- sous-corpus [LAROUSSE] (*Grand Universel Larousse*-GUL et *Grand Larousse Informatisé*-GLI)

Caractérisation : textes de type encyclopédique

Une annotation manuelle des segments d'obsolescence

- annotateurs de chez Larousse et moi-même
- directive d'annotation peu guidée

Quelles unités d'analyse et de traitement ?

Constats issus de l'annotation manuelle :

- les segments à mettre à jour sont de tailles variables : du syntagme à la section entière
  $\Rightarrow$ introduction des notions de **segment minimal** et de **cadre d'interprétation**
- co-existance de deux types d'informations évolutives :
  - les **réadaptations** : l'information est devenue fausse
  - les **réactualisations** : l'information reste vraie mais n'est plus pertinente

$\Rightarrow$ l'obsolescence est un phénomène complexe qu'il faut simplifier.

$\Rightarrow$ parce qu'on vise un traitement automatique, il faut définir une unité de traitement fixe/stable.

Contraintes

On veut pouvoir traiter :

- des unités linguistiques de petite taille : adverbiaux, temps verbaux, etc.
- des adverbiaux spécifiques de type introducteurs de cadres
- les unités "titres" en tant qu'élément de structure mais aussi sémantique
- les positions des éléments dans le document

Quelles unités ?

Dans la réalité des textes, les segments obsolescents sont :

des segments de taille variable : mots, groupes de mots, phrases, groupe de phrase

mots, groupes de mots

Actualités

Établir une liste exhaustive des avancées récentes de la recherche médicale est impossible tant les progrès sont nombreux.

Toutefois, il convient de rappeler un certain nombre de découvertes très récentes.

En 2003, l'une des grandes priorités de la recherche médicale internationale a concerné le sida.

Quelles unités ?

Dans la réalité des textes, les segments obsolescents sont :

des segments de taille variable : mots, groupes de mots, phrases, groupe de phrase

phrases, titres

Actualités

Établir une liste exhaustive des avancées récentes de la recherche médicale est impossible tant les progrès sont nombreux.

Toutefois, il convient de rappeler un certain nombre de découvertes très récentes.

En 2003, l'une des grandes priorités de la recherche médicale internationale a concerné le sida.

Quelles unités ?

Dans la réalité des textes, les segments obsolescents sont :

des segments de taille variable : mots, groupes de mots, phrases, groupe de phrase

paragraphes, titres

Actualités

Établir une liste exhaustive des avancées récentes de la recherche médicale est impossible tant les progrès sont nombreux.

Toutefois, il convient de rappeler un certain nombre de découvertes très récentes.

En 2003, l'une des grandes priorités de la recherche médicale internationale a concerné le sida.

Quelles unités ?

Dans la réalité des textes, les segments obsolescents sont :

des segments de taille variable : mots, groupes de mots, phrases, groupe de phrase

sections titrées

Actualités

Établir une liste exhaustive des avancées récentes de la recherche médicale est impossible tant les progrès sont nombreux.

Toutefois, il convient de rappeler un certain nombre de découvertes très récentes.

En 2003, l'une des grandes priorités de la recherche médicale internationale a concerné le sida.

Quelles unités ?

Dans la réalité des textes, les segments obsolescents sont :

des segments de taille variable : mots, groupes de mots, phrases, groupe de phrase

phrases, titres

Actualités

Établir une liste exhaustive des avancées récentes de la recherche médicale est impossible tant les progrès sont nombreux.

Toutefois, il convient de rappeler un certain nombre de découvertes très récentes.

En 2003, l'une des grandes priorités de la recherche médicale internationale a concerné le sida.

Choix final

$\Leftrightarrow$ une unité linguistique consensuelle, un segment visuel fort et une unité facilement repérable automatiquement

Résultats de l'annotation manuelle : l'obsolescence, un phénomène rare et complexe

Proportion de segments obsolescents dans le corpus

nombre total de phrases	9916
nombre de phrases obsolescentes	1508
pourcentage de phrases obsolescentes	**15.2 %**

Accords de jugement sur l'obsolescence (*Coefficient r de Finn*)

$\Rightarrow$ scores entre 0.75 et 0.83

- des scores acceptables : l'automatisation de la tâche est envisageable
- mais qui restent peu élevés : les performances de la machine sont à relativiser

Comprendre l'obsolescence

Des indices sémantiques fréquents dans les segments d'obsolescence

- des dates $>$ 2003 : « *prévu en 2012* »
- des adverbes temporels déictiques : « *aujourd'hui* »
- des valeurs chiffrées particulières dans des rubriques spécifiques : le nombre d'habitants en géographie *vs.* en histoire

Mais pas de marqueur explicite de l'obsolescence

$\longrightarrow$ recherche des combinaisons d'indices qui favorisent ou non l'interprétation obsolescente d'un segment textuel

Plan

Choix des indices : méthodologie

Observation du corpus annoté

relevé d'expressions récurrentes dans les segments d'obsolescence et classification

Mon expérience de linguiste

- prise en compte de la multiplicité des indices (Biber [1988, 1989], HoDac [2007], Widlöcher [2008], projet ANNODIS)
- prise en compte de la structuration discursive [Teufel, 1999, Bouffier, 2008] :
  - les titres [Jacques and Rebeyrolle, 2006], les positions textuelles [Mani, 2001]
  - l'hypothèse de l'encadrement du discours (Charolles [1997], projet GEOSEM)

Choix des indices : méthodologie

- mise en évidence d'une très **vaste palette d'indices linguistiques** potentiellement pertinents pour l'obsolescence : on ne sait pas si les indices sélectionnés seront réellement pertinents / des hypothèses sur leur pertinence
- recherche de la **validation d'un système d'indices** : les combinaisons d'indices sont-elles pertinentes pour notre tâche ?

Des indices linguistiques variés ...

Des indices intra-phrastiques

- des **expressions temporelles** : « *aujourd'hui* », « *dans les années à venir* »
- des **informations aspectuelles** : « *être en train de* », « *être en cours de* »
- des **entités nommées** : mesures (« *30 hab./km^2* »), lieux (« *à Paris* »), sigles (« *UNESCO* »)
- des **informations modales** : types de phrases (assertion, exclamation, interrogation), « *il est urgent de* », « *Selon le rapport de l'INSEE* », « *malheureusement* »

... et à granularité variable

Des indices discursifs

- **indices positionnels phrastiques** : la position des indices intra-phrastiques dans les phrases
- **indices positionnels textuels** : la position des phrases dans le paragraphe, du paragraphe dans la section
- **indices hiérarchiques** : les titres
- **indices externes** : le thème du texte fourni par l'encyclopédie

Des indices linguistiques variés

indices de temps

Actualités

Établir une liste exhaustive des avancées récentes de la recherche médicale est impossible tant les progrès sont nombreux.

Toutefois, il convient de rappeler un certain nombre de découvertes très récentes.

En 2003, l'une des grandes priorités de la recherche médicale internationale a concerné le sida.

Des indices linguistiques variés

indices de temps

argumentatifs

Actualités

Établir une liste exhaustive des avancées récentes de la recherche médicale est impossible tant les progrès sont nombreux.

Toutefois, il convient de rappeler un certain nombre de découvertes très récentes.

En 2003, l'une des grandes priorités de la recherche médicale internationale a concerné le sida.

Des indices linguistiques variés

indices de temps
argumentatifs
indices « modaux »

Actualités

Établir une liste exhaustive des avancées récentes de la recherche médicale est impossible tant les progrès sont nombreux.

Toutefois, il convient de rappeler un certain nombre de découvertes très récentes.

En 2003, l'une des grandes priorités de la recherche médicale internationale a concerné le sida.

Des indices à granularité variable

indices de temps

argumentatifs

indices « modaux »

titres et phrases
(indications positionnelles)

Actualités

Établir une liste exhaustive des avancées récentes de la recherche médicale est impossible tant les progrès sont nombreux.

Toutefois, il convient de rappeler un certain nombre de découvertes très récentes.

En 2003, l'une des grandes priorités de la recherche médicale internationale a concerné le sida.

Des indices à granularité variable

indices de temps

argumentatifs

indices « modaux »

titres et phrases
(indications positionnelles)

paragraphe et section titrée
(indications positionnelles)

Actualités

Établir une liste exhaustive des avancées récentes de la recherche médicale est impossible tant les progrès sont nombreux.

Toutefois, il convient de rappeler un certain nombre de découvertes très récentes.

En 2003, l'une des grandes priorités de la recherche médicale internationale a concerné le sida.

Les segments d'obsolescence, des unités complexes

Nous voulons exploiter

- une **grande variété d'indices linguistiques** (classes génériques + sous-catégorisation)
- des **niveaux d'analyse à granularité variable**

Hypothèse sur les possibles réalisations des marqueurs de l'obsolescence

- un indice lexical univoque
- une combinaison d'indices de même type et de même granularité
- une combinaison d'indices différents

Plan

Repérage et annotation automatique des indices présentés

Processus entièrement automatique

$\Leftrightarrow$ plateforme LinguaStream [Widlöcher and Bilhaut, 2005] : environnement de développement dédié aux traitements T.A.L. qui facilite la création de :

- segmenteur (en mots, en phrases)
- lexiques
- grammaires ProLog
- Macro Expressions Régulières
- ...

Développement d'un ensemble de traitements spécifiques (outil ALIDIS)

$\Rightarrow$ modules de traitement du temps, de la modalité, de l'aspect, des positions,...

Un exemple

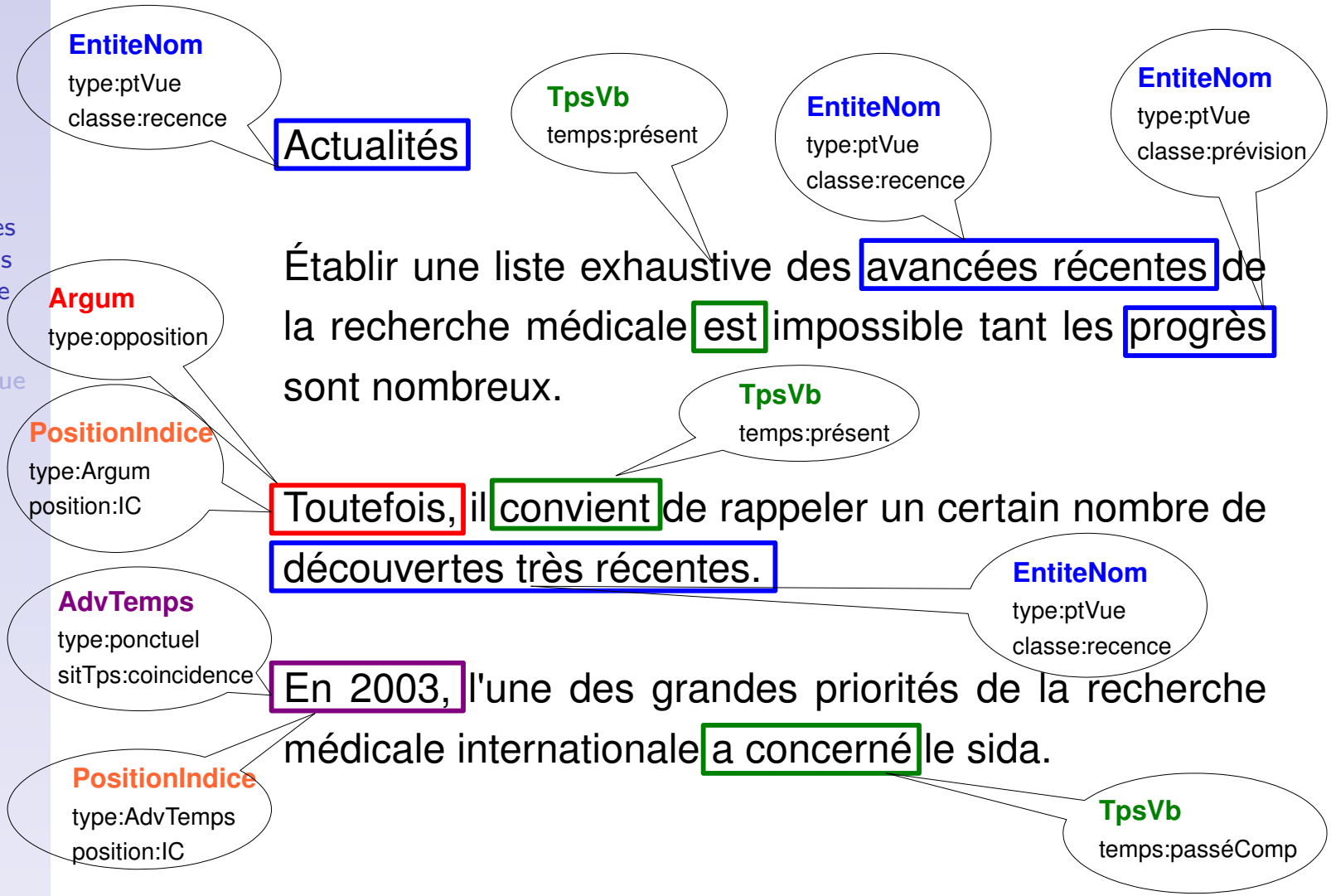
Actualités

Établir une liste exhaustive des avancées récentes de la recherche médicale est impossible tant les progrès sont nombreux.

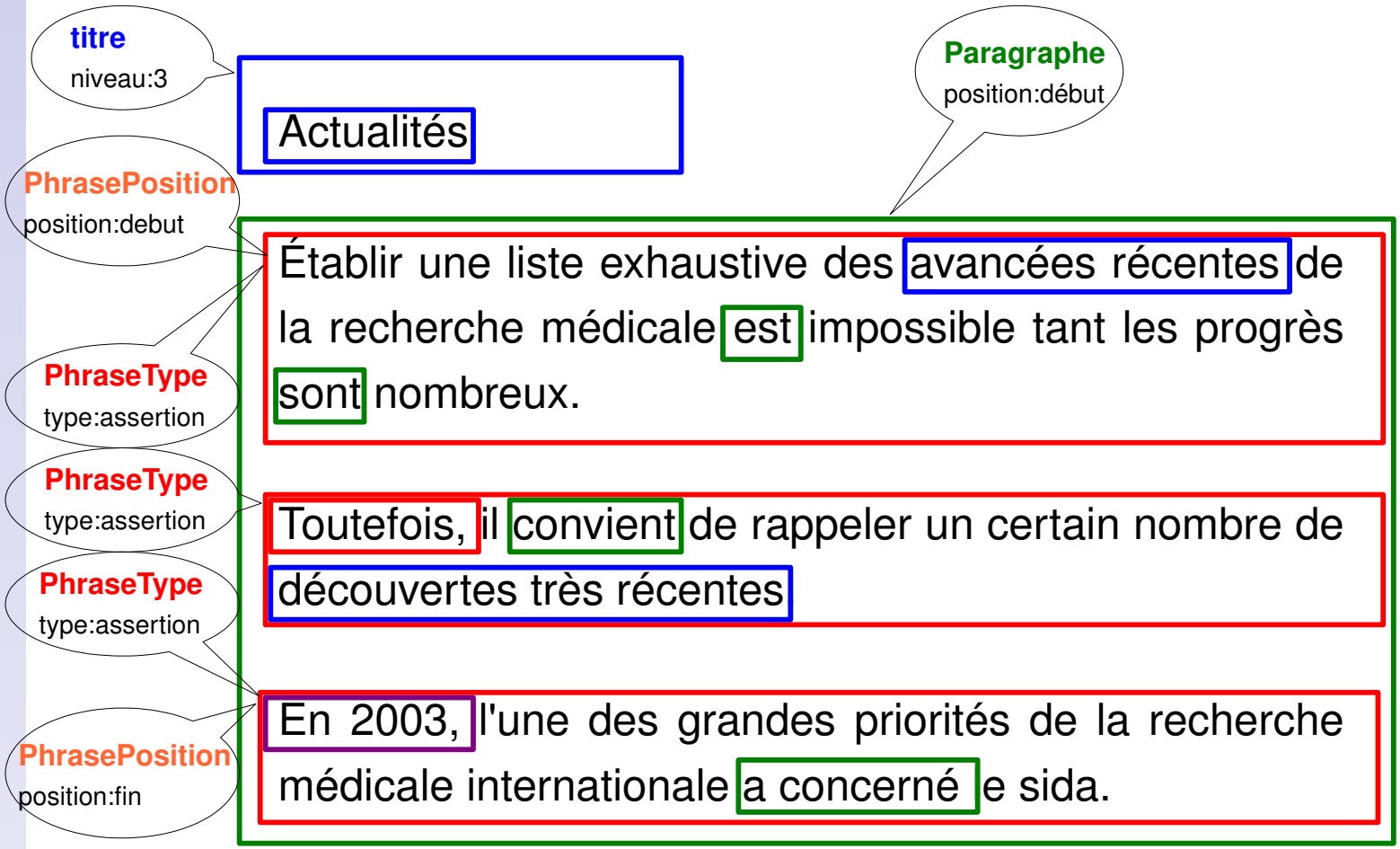
Toutefois, il convient de rappeler un certain nombre de découvertes très récentes.

En 2003, l'une des grandes priorités de la recherche médicale internationale a concerné le sida.

Annotation des indices intraphrastiques



Annotation des indices discursifs



Performances de l'outil ALIDIS

Évaluation

$\Leftrightarrow$ menée à la main sur $1/10^e$ du corpus

précision : 93 %

rappel : 85 %

$\Rightarrow$ les repérages sont suffisamment fiables pour envisager une exploitation automatique et à grande échelle

Notre objectif : repérer les combinaisons d'indices pertinentes pour l'obsolescence

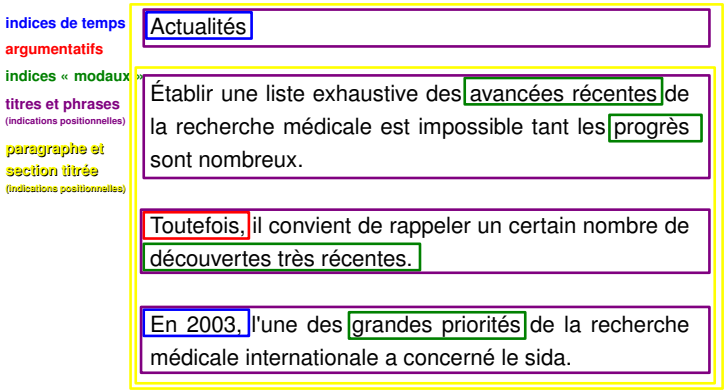
$\Rightarrow$ on pourrait le faire à la main ... mais ...

- environ 10 000 phrases
- environ 150 indices

$\Rightarrow$ on se tourne vers les statistiques
$\Leftrightarrow$ il faut transformer des données textuelles complexes en matrice à deux dimensions

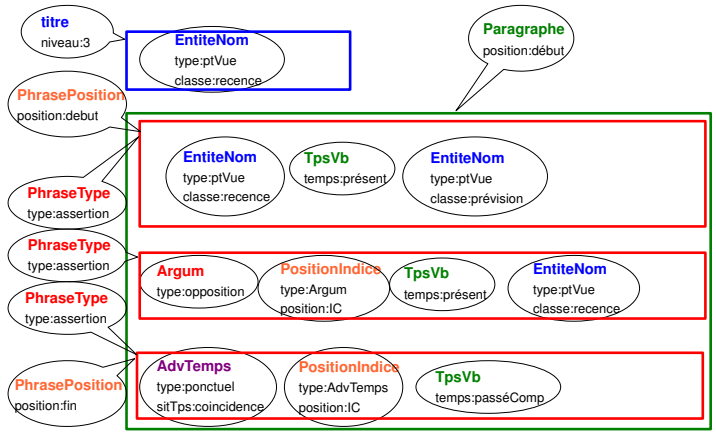
Transformer des données textuelles complexes en matrice à deux dimensions

Étape 1 (outil ALIDIS)



⇒

Étape 2 (outils ALIDIS et OCAS)



⇓???

Objectif (entrée pour l'outil STAAT)

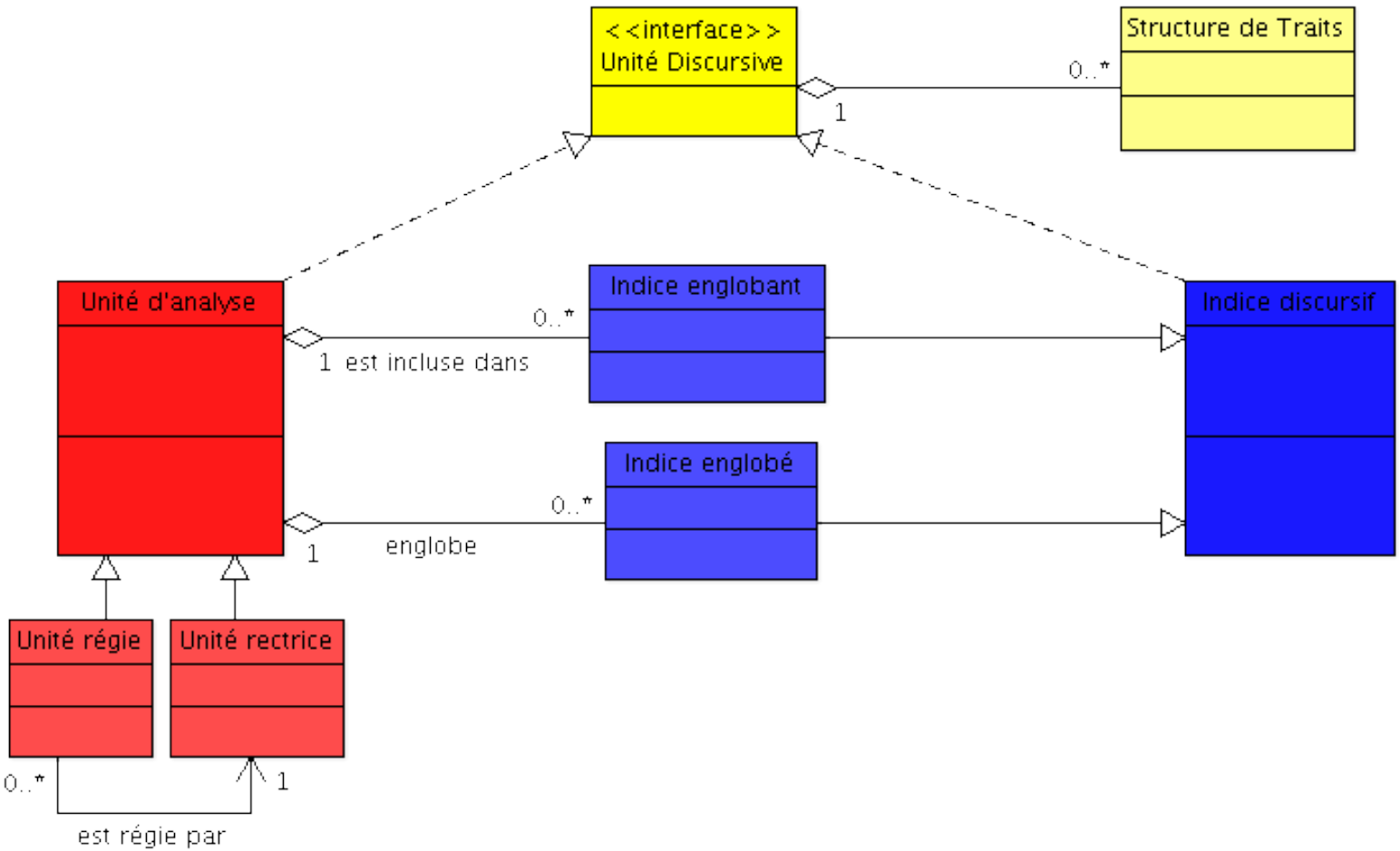
individu	var_1	var_2	var_3	var_4	var_x
$individu_1$	a	b	c	d	e
$individu_2$	a	b	c	d	e
$individu_3$	a	b	c	d	e
$individu_n$	a	b	c	d	e

Contraintes de représentation

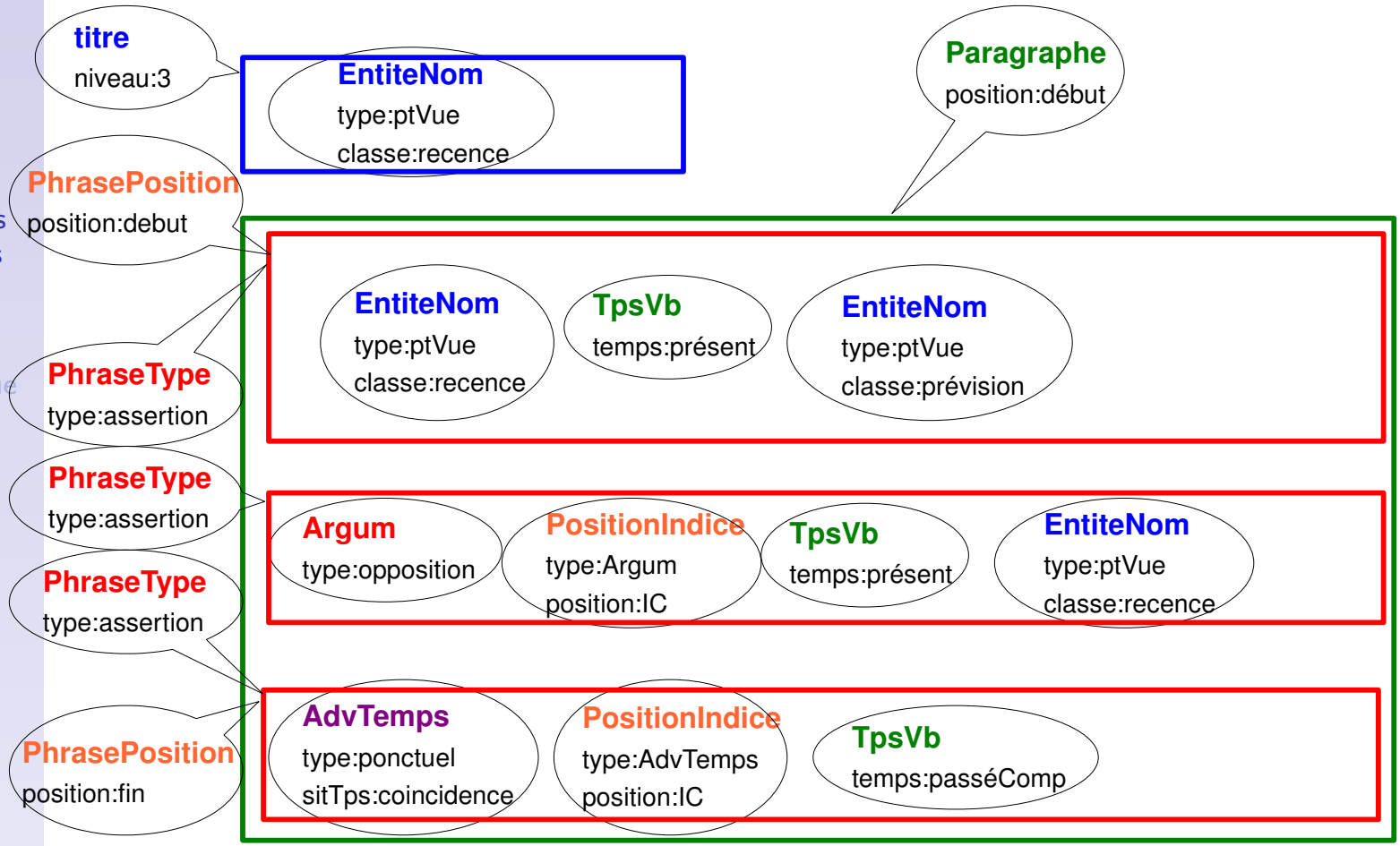
- **Contrainte 1 :** un titre et une phrase doivent pouvoir être décrits à travers n'importe quel type d'indice
- **Contrainte 2 :** on ne connait pas *a priori* le nombre d'indices présent à l'intérieur de l'unité phrase/titre ni le niveau de profondeur de l'unité
- **Contrainte 3 :** une phrase peut être obsolescente mais un titre ne le sera jamais, il est un prédicteur potentiel de l'obsolescence (*cf.* notion d'héritage de contexte, Zerida et al. [2006])
- **Contrainte 4 :** on ne connait pas *a priori* le format d'entrée pour les statistiques ; il faut un stockage des données au plus proche de la réalité des textes
- **Contrainte 5 :** on a besoin de pouvoir modifier les indices et leurs annotations sémantiques sans avoir à tout réécrire

$\Rightarrow$ pour gérer cette complexité, mise en place d'un modèle de représentation des indices discursifs pour leur traitement automatique (statistique)

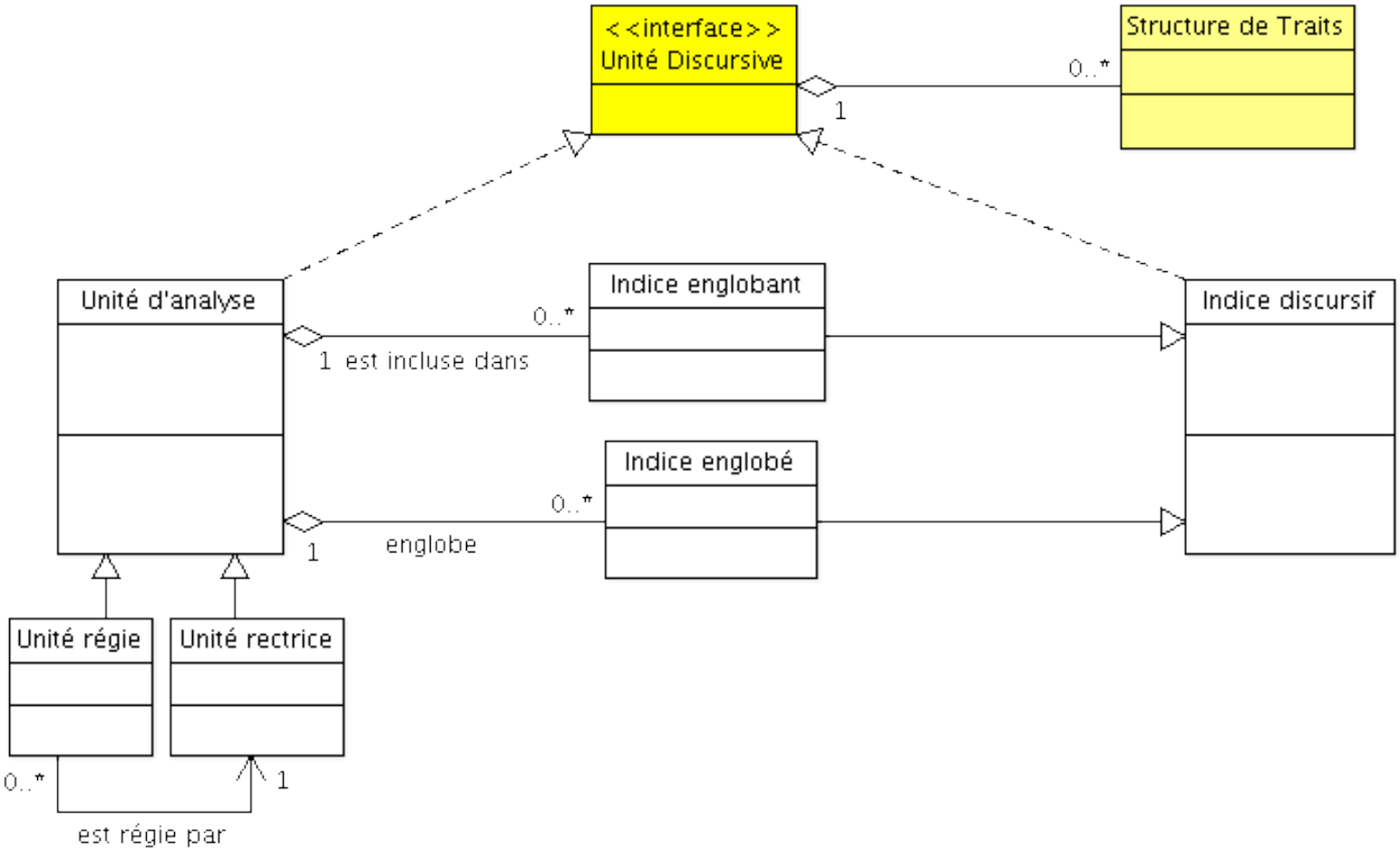
Un modèle UML de représentation des indices de discours : gérer la variabilité du grain d'analyse



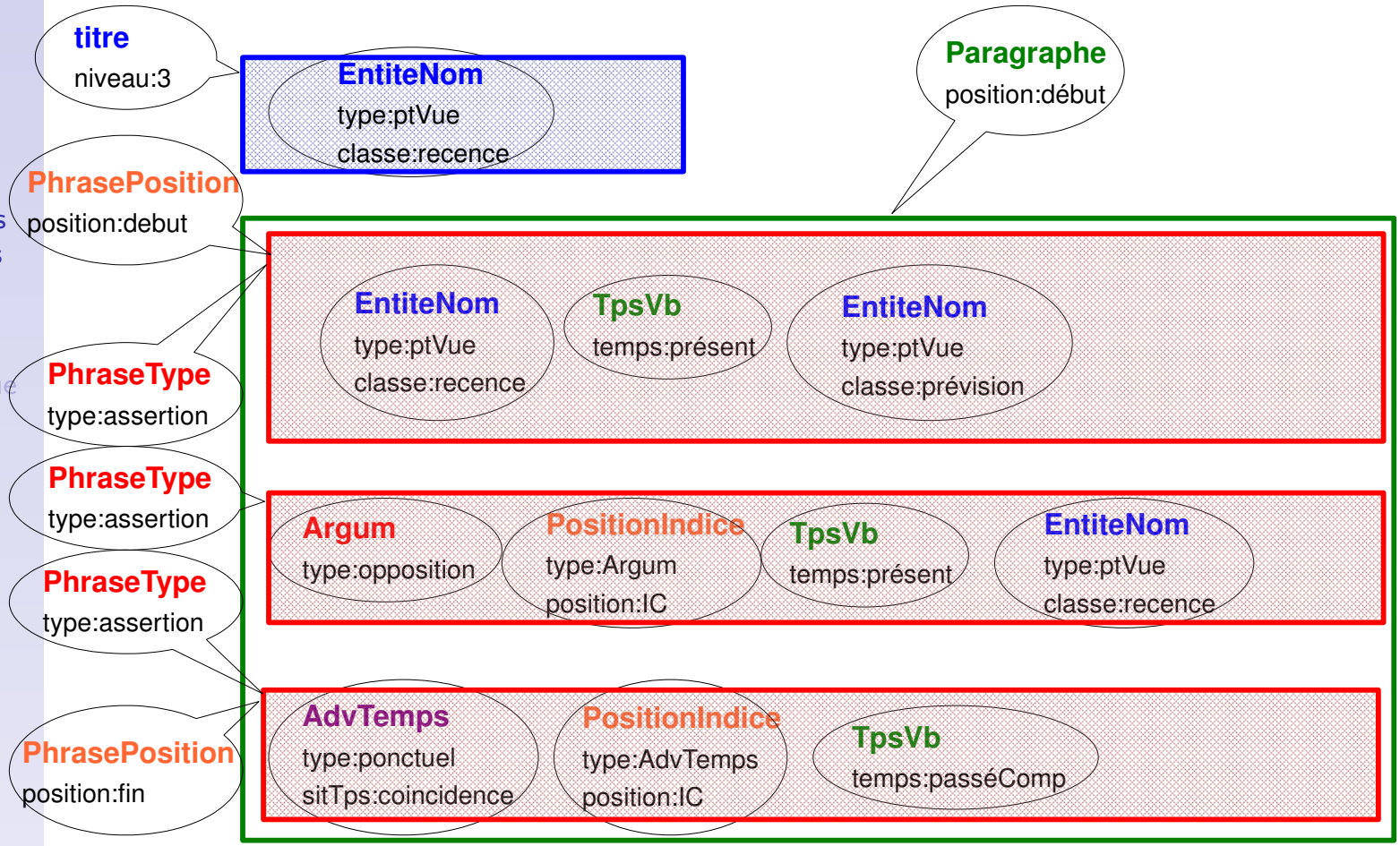
Tout est **unité discursive**



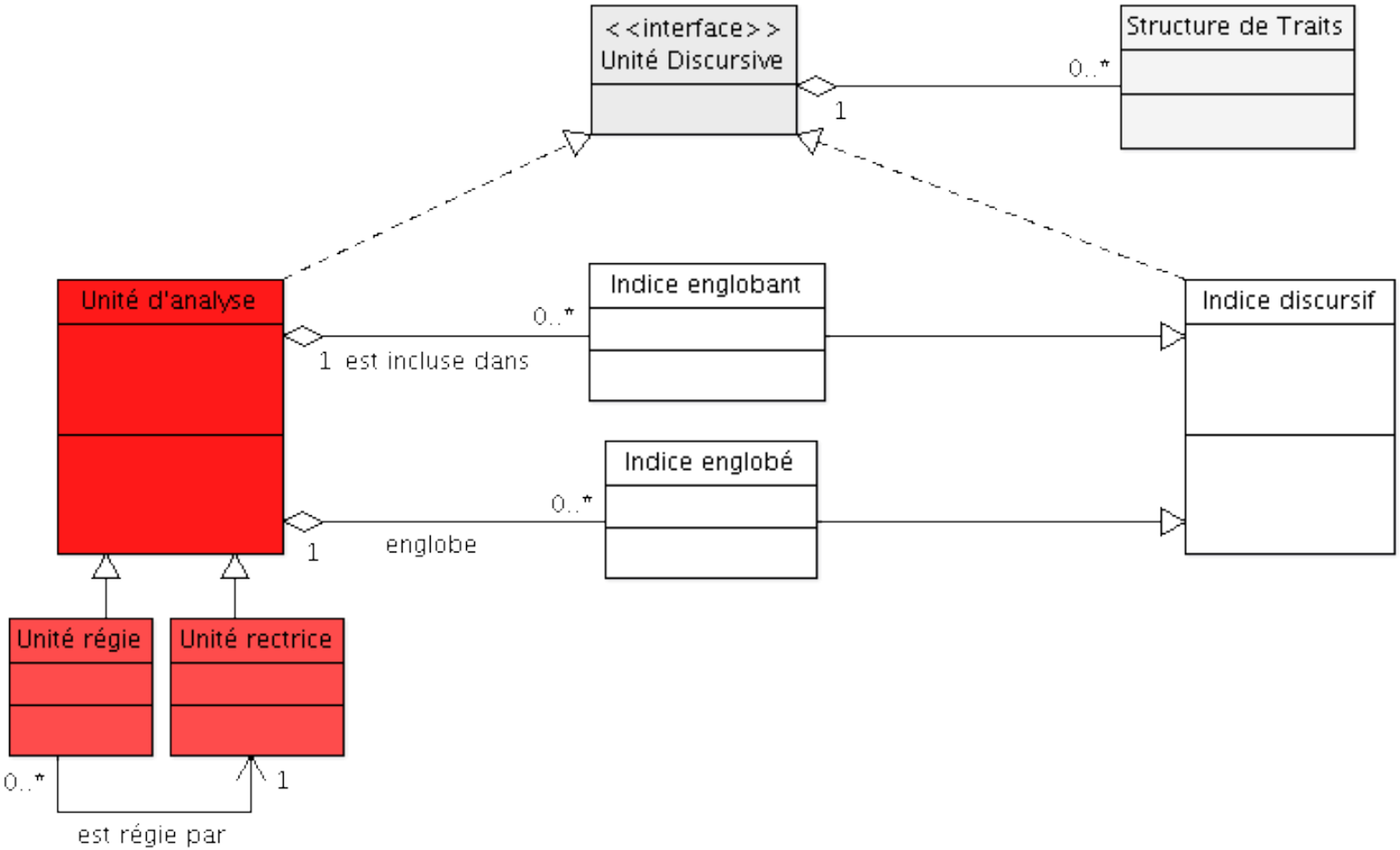
Les unités discursives



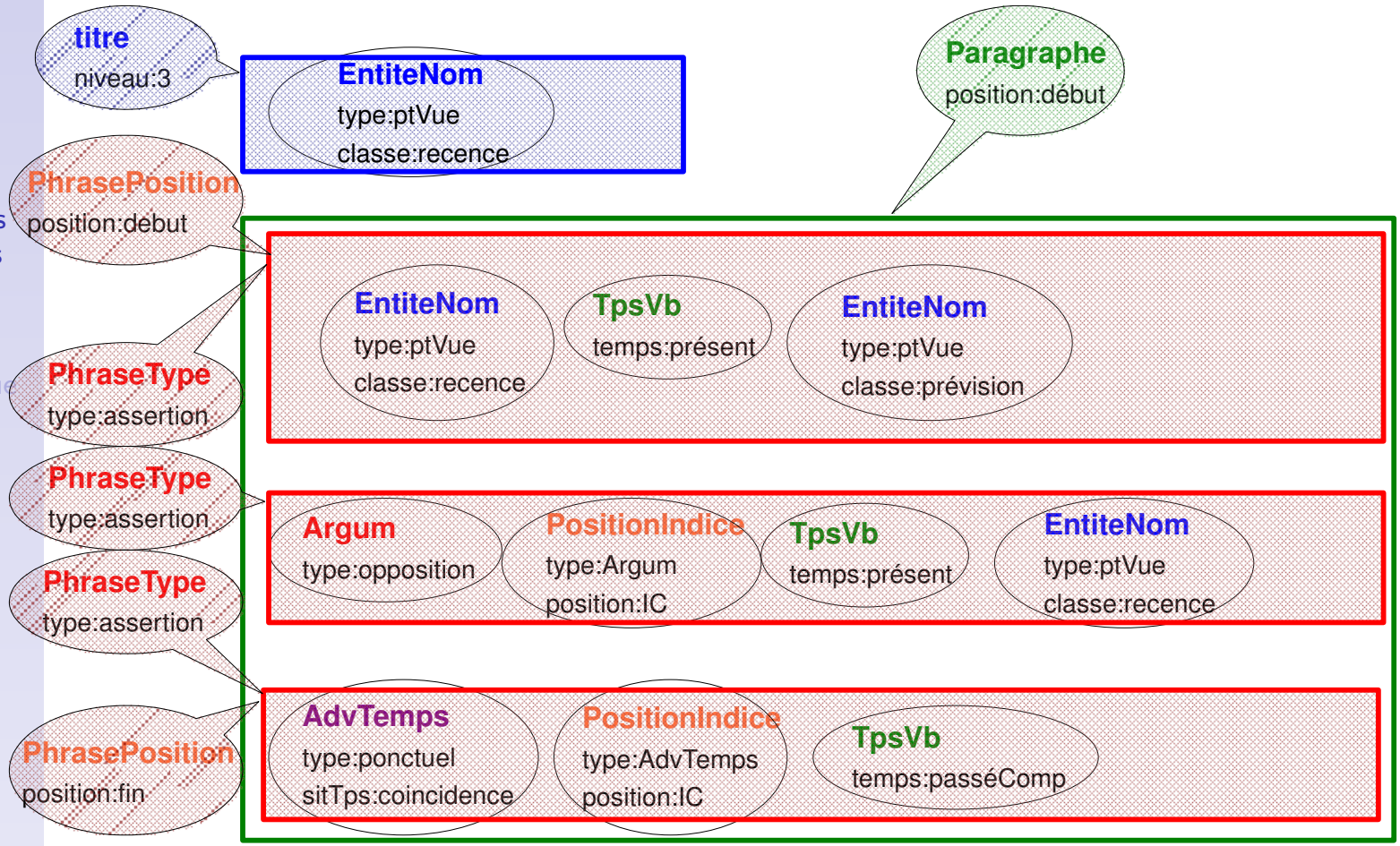
Choix des unités d'analyse



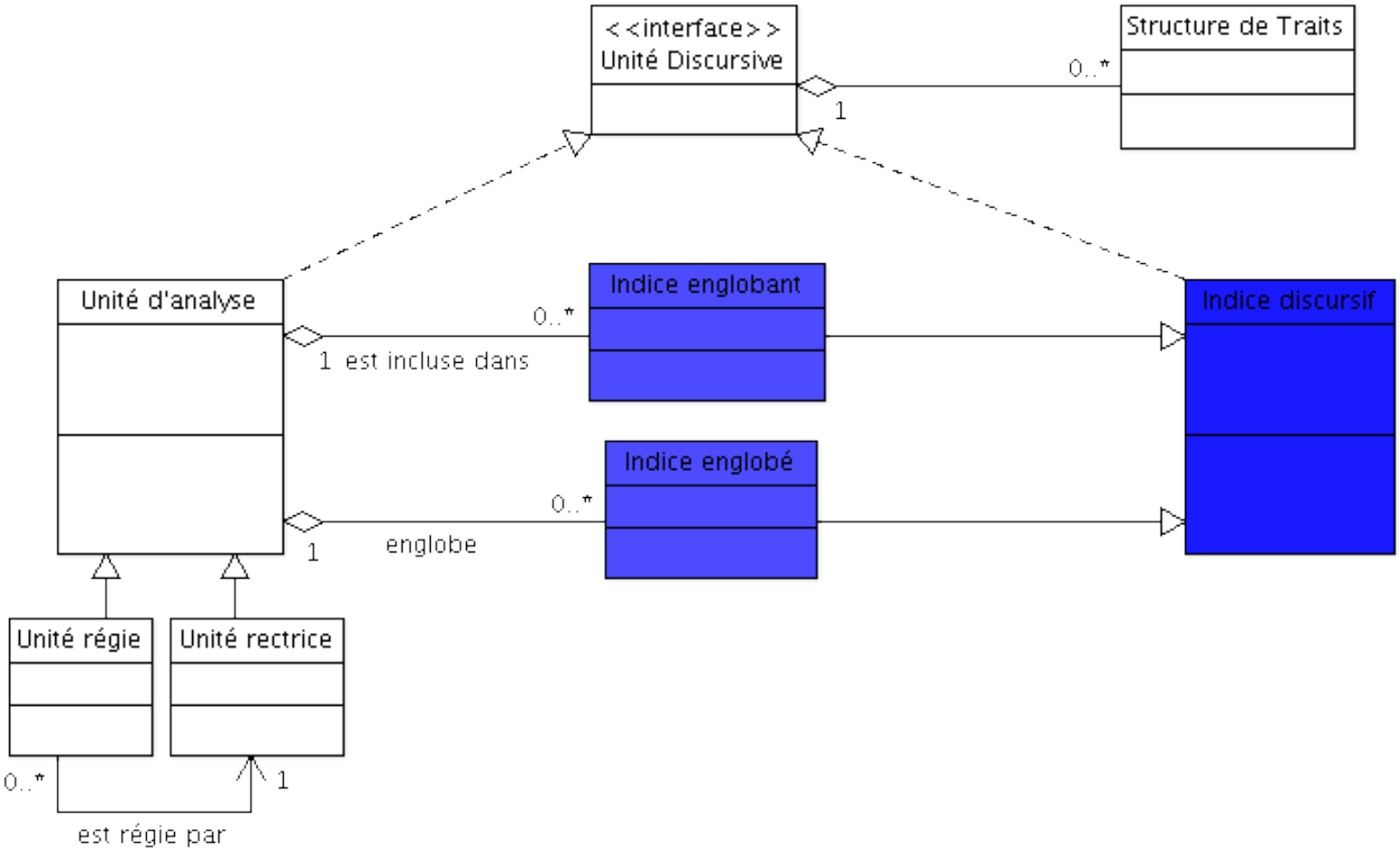
Choix des unités d'analyse



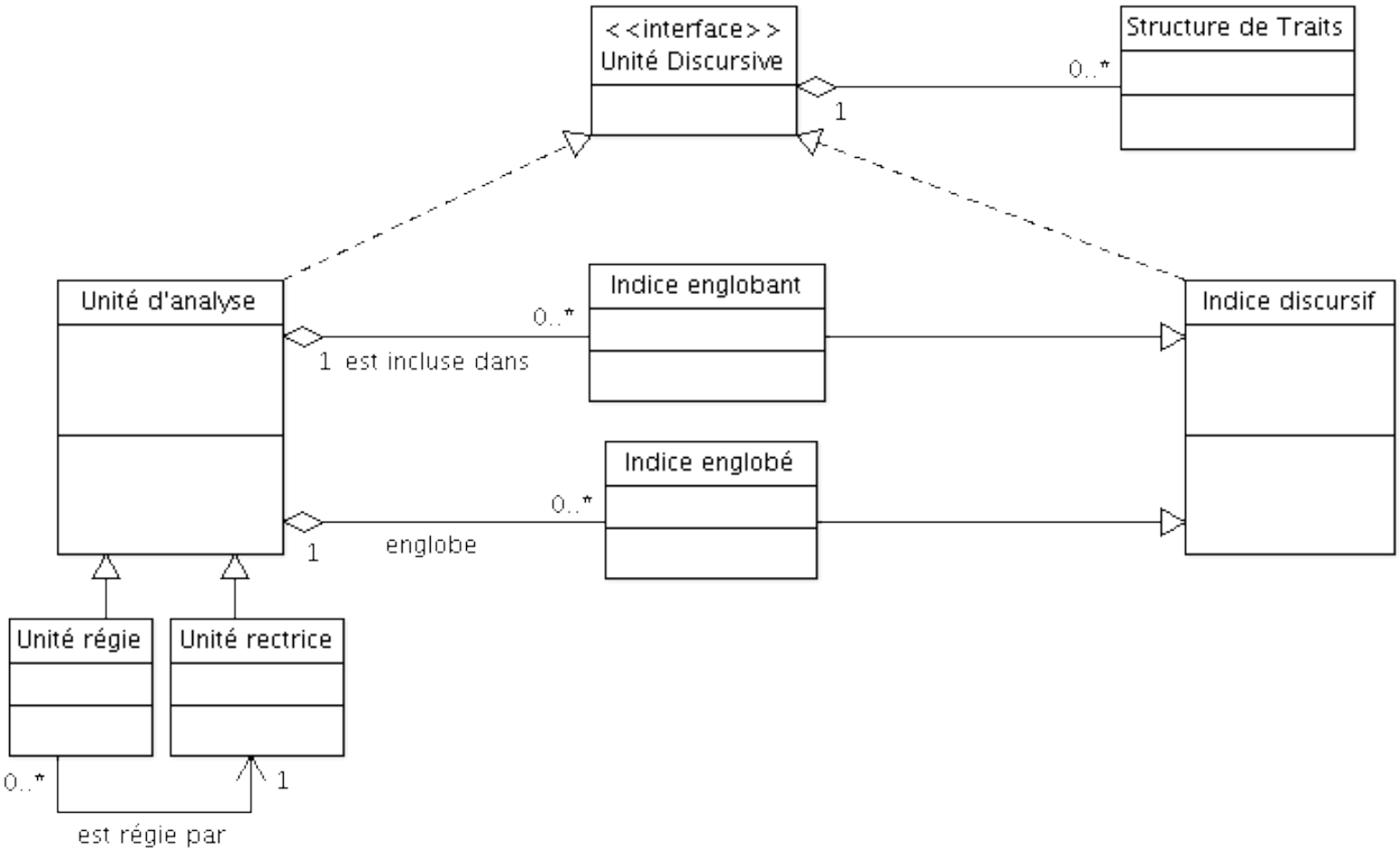
Observation des relations des UA avec les ID



Observer les indices discursifs



Le modèle UML de représentation des indices de discours comme modèle de création de la base de données



Représentation en 2 dimensions : variables (indices) possibles pour un individu donné (phrase)

Nom des variables	**Valeur**
indices externes	
encyclopedie.type :GUL	{0,1}
zone.rubriqueName :ArtLitt	{0,1}
indices intra-phrastiques	
exprTemps.nature :deictique ;sitTps :coincidence	{0,1}
exprTemps.nature :anaphorique ;sitTps :indetermine	{0,1}
entiteNom.classe :geopolitique	{0,1}
entiteNom.classe :personne	{0,1}
argum.relation :precision	{0,1}
tpsVbx.temps :futur	{0,1}
tpsVbx.temps :présent	{0,1}
ptVue.type :prevision	{0,1}
ptVue.type :distance	{0,1}
periVbs.accomplissement :deroulement	{0,1}
PhraseType.type :assertion	{0,1}
indices positionnels phrastiques	
position.typeIndice :exprTemp ;typePos :IC	{0,1}
indices hiérarchiques	
titre−>advTemps.nature :deictique ;sitTps :coincidence	{0,1}
titre−>entiteNom.classe :personne	{0,1}
titre−>niveau :3	{0,1}
indices positionnels textuels	
PhrasePosition.position :debutParagraphe	{0,1}
premierParag.position :debutDivisionSeul	{0,1}
dernierParag.position :finZone	{0,1}

Plan

Pourquoi des statistiques ?

Un outil pour

- traiter des données conséquentes (pour nous, matrice de 9916 individus et 150 variables)
- prouver/infirmer des hypothèses de recherche
- faire émerger des informations nouvelles

$\Rightarrow$ des statistiques descriptives
$\Rightarrow$ des statistiques prédictives

Corrélation de la variable *obsol* avec toutes les autres : statistiques de base

Mesure de la pertinence des indices

- **indices corrélés positivement :** par exemple, les entités nommées de type *géopolitique* ou de type *mesure évolutive* ou encore les adverbiaux temporels de type *déictique coïncidence*
- **indices corrélés négativement :** par exemple, les entités nommées de type *personne* ou encore les adverbiaux temporels de type *ponctuel antériorité++*
- **indices pas corrélés :** par exemple, le futur

Corrélation de la variable *obsol* avec des combinaisons de variables : Analyse en Composantes Principales

Description

- **du corpus** : des oppositions fortes entre les sous-corpus (ATLAS vs. LAROUSSE et GLI vs. GUL)
- **des segments d'obsolescence** : des regroupements intéressants (par exemple indices de temps et de mesure)

Mais surtout

$\Rightarrow$ l'obsolescence est un problème complexe, qui ne peut être résolu simplement (pas d'indice unique, pas de combinaisons d'indices fortes). Il faut combiner les indices si on veut repérer l'obsolescence.

Apprentissage automatique

Objectif de l'AA : exploitation de méthodes statistiques sur des données annotées qui permet de prendre de nouvelles décisions sur des textes nouveaux

Intérêt : démarche exploratoire visant à déterminer les indices/combinaisons d'indices pertinentes dans les segments d'obsolescence

Besoin spécifique : un modèle interprétable

Technique de classification supervisée

- valide l'intérêt et la pertinence des indices dans les segments d'obsolescence
- vérifie si une machine peut repérer automatiquement ces segments sur la base des indices présentés

$\Rightarrow$ Utilisation d'un système à base de **règles d'association** (Rioult et al. [2008])

Les règles d'association

Format des règles

Si A, alors B

- A et B sont des conjonctions d'attributs
- A est la condition, la prémisse de la règle
- B est la conclusion (sur une valeur de classe).

Exemple de règle d'association

$exprTemp.nature : deictique; sitTps : coincidence \wedge entiteNom.classe : mesure; sousClasse : evolutif \longrightarrow classe : obsol$

Résultats

- Les règles sont nombreuses : environ 1500, redondantes
- chaque règle couvre au moins 9 phrases

Évaluation quantitative

Des résultats encourageants

- évaluation du système (validation croisée) :

  rappel 78 %
  précision 34 %

- évaluation du prototype d'aide à la mise à jour de texte par les experts (comparaison machine/homme) :

  rappel 65 %
  précision 37 %
  cœfficient r de Finn 0.70 (accord entre humains entre 0.75 et 0.83)

$\Rightarrow$ il vaut mieux favoriser le rappel que la précision

Toutes les associations d'indices sont représentées

indices positionnels phrastiques + indices intra-phrastiques

fin de paragraphe + verbe au conditionnel + entité nommée de type *mesure évolutive*

§ *[...] L'Union européenne à elle seule se* **serait** *dépossédée d'un patrimoine de* **215 milliards de dollars**.

plusieurs indices intra-phrastiques

expression temporelle *déictique, coïncidence* + entité nommée de type *mesure évolutive*

§ *Les Noirs représentent* **aujourd'hui 12 %** *de la population ; plus de* **50 %** *d'entre eux sont* **encore** *concentrés dans le Sud historique.*

Les indices de type titre et les indices positionnels

mesurer l'impact des différents indices

Pour cela, nous avons créé cinq vues différentes du corpus [ENCYCLO] :

- *corpusComplet* : une vue qui prend en compte tous les indices ;
- *corpusIPseuls* : une vue qui prend en compte uniquement les indices intra-phrastiques ;
- *corpusIPHierar* : une vue qui prend en compte les indices intra-phrastiques et les indices hiérarchiques ;
- *corpusIPPos* : une vue qui prend en compte les indices intra-phrastiques et les indices positionnels ;
- *corpusEpure* : un corpus « épuré » dans lequel sont enlevées les variables non significatives (en fonction des résultats des statistiques de base, DESCO, et de l'ACP).

Résultats

	Précision	Rappel	F-Score
corpusIPseuls	38	37	37.5
corpusIPHierar	39.9	45.6	42.5
corpusIPPos	33.2	56.7	41.9
corpusComplet	32.9	**78.8**	46.4
corpusEpure	**38.7**	**62.3**	47.7

Interprétation - Discussion

- exploiter les indices intra-phrastiques uniquement est insuffisant
- gain mesurable avec l'exploitation d'indices de type hiérarchique (+ 5 sur le F-score), ou positionnel (+ 4.4) ou les deux (+ 8.9)
- les indices hiérarchiques favorisent la précision
- les indices positionnels privilégient le rappel
- il vaut mieux ne pas considérer trop d'indices : le corpus épuré contient tous les types d'indices qui sont significatifs

Interprétation - Discussion : les indices positionnels

Position de la phrase dans le paragraphe

- les premières phrases de paragraphe associées à des indices temporels de type *coïncidence* ou à des entités nommées de type *mesure*.
- les dernières phrases de paragraphe associées au conditionnel et à des entités nommées de type *géopolitique*.

$\Rightarrow$ *cf.* HoDac [2007] = importance des éléments en première phrase de paragraphe ou de section sur la variation textuelle

Interprétation - Discussion : les indices positionnels

Position du paragraphe dans le document

- la position d'introduction de section est corrélée à des indices temporels ou des entités nommées de type *géopolitique* $\longrightarrow$ un paragraphe introductif a tendance à situer le décor temporel ou référentiel (*i.e.* de quoi on va parler).
- la position conclusive est associée aux entités nommées de type *mesure* $\longrightarrow$ un paragraphe conclusif se base sur des exemples précis, concrets qui font appel à des valeurs chiffrées précises.

$\Rightarrow$ *cf.* Marcu [2000] = relation entre la présence de marques linguistiques particulières et leur apparition dans des positions paragraphiques précises pour juger automatiquement de l'*importance* d'une phrase (RA).

Interprétation - Discussion : les indices hiérarchiques

Les indices dans les titres

- une entité nommée de type *géopolitique, mesure ou lieu* ou une expression temporelle de type *déictique* dans un titre est significativement corrélée à l'obsolescence
- une faible présence d'indices de rhétorique (*i.e.* de connecteurs discursifs) et d'indice de nouveauté (*i.e.* de point de vue) dans les titres (*cf.* Ibekwe-SanJuan [2005])

$\Rightarrow$ situation du décor temporel ou référentiel

Interprétation - Discussion : les rubriques thématiques

caractérisation du texte en fonction des rubriques thématiques

certains indices ou combinaisons d'indices sont pertinents pour une rubrique particulière.
la combinaison « *entité nommée lieu* + *entité nommée mesure* » est fortement associée à l'obsolescence dans un texte géographique mais est contre-productive en histoire

$\Rightarrow$ *cf.* Zerida et al. [2006] : différence significative dans l'organisation de l'écrit et dans le style de trois types de textes biomédicaux (articles de recherche, de synthèse, cliniques,...).

Plan

Conclusion I

Remarques générales sur l'obsolescence

- une réponse appropriée à la question de la mise à jour dans les documents encyclopédiques
- un phénomène rare et complexe
- qui peut être appréhendé par des outils linguistiques

Mise en œuvre d'un outil linguistique

- fondé sur la notion de marqueur discursif
- qui suppose l'intérêt des combinaisons d'indices hétérogènes

Conclusion II

Recherche exploratoire sur la validité d'un système de combinaisons d'indices

- un système basé sur la multiplicité des indices est une réponse adéquate
- l'ensemble des types d'indices pris en compte sont nécessaires
- mise en évidence de l'importance de la structure du discours : position des segments et titres sont centraux dans notre approche
  $\Rightarrow$ ne sont-ils que des outils ou au contraire des indices linguistiques à part entière ?

Conclusion III

Méthodologie mise en œuvre

- évolutive (programmes et ressources)
- reproductible en partie (sous réserve des corpus sous licence)
- réutilisable

Bibliographie

D. Biber. A typology of english texts. *Linguistics*, 27 :3–43, 1989.

D. Biber. *Variation across speech and writing*. Cambridge University Press, Cambridge, 1988.

A. Bouffier. *Analyse discursive automatique de textes - Application à la modélisation de textes incitatifs*. PhD thesis, Université Paris Nord - Villetaneuse, 2008.

M. Charolles. L'encadrement du discours, univers, champs, domaine et espaces. *Cahiers de Recherche linguistique*, 6, 1997.

M. HoDac. *La position initiale dans l'organisation du discours : une exploration en corpus*. PhD thesis, Université de Toulouse 2 - Le Mirail, 2007.

F. Ibekwe-SanJuan. Annotation d'indices de nouveautés dans les écrits scientifiques et techniques. In *Colloque Indice, Index, Indexation*, 2005.

M.-P. Jacques and J. Rebeyrolle. Titres et structuration des documents. In *Actes du Colloque International Discours et Document*, pages 1–12, Caen, France, 2006.

I. Mani. *Automatic summarization*. John Benjamins Publishing Compagny, Amsterdam/Philadelphie, 2001.

D. Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. The MIT Press, 2000.

F. Rioult, B. Zanuttini, and B. Crémilleux. Apport de la négation pour la classification supervisée à l'aide d'associations. In *Conférence d'Apprentissage*, pages 183–196, 2008.

S. Teufel. *Argumentative Zoning*. PhD thesis, Université de Edimbourg, 1999.

A. Widlöcher. *Analyse macro-sémantique des structures rhétoriques du discours. Cadre théorique et modèle opératoire.* PhD thesis, Université de Caen, 2008.

A. Widlöcher and F. Bilhaut. La plate-forme linguastream : un outil d'exploration linguistique sur corpus. In *Actes de la 12e Conférence Traitement Automatique du Langage Naturel (TALN)*, Dourdan, France, 2005.

N. Zerida, N. Lucas, and B. Crémilleux. Combinaison de descripteurs linguistiques et de structure pour la fouille d'articles biomédicaux. In *Schedae*, pages 69–78, 2006.