

# Acquisition de liens sémantiques à partir d'éléments de mise en forme des textes : exploitation des structures énumératives

Jean-Philippe Fauconnier  
encadrantes : Mouna Kamel et Nathalie Aussenac-Gilles



Institut de Recherche en Informatique de Toulouse  
Équipe MELODI

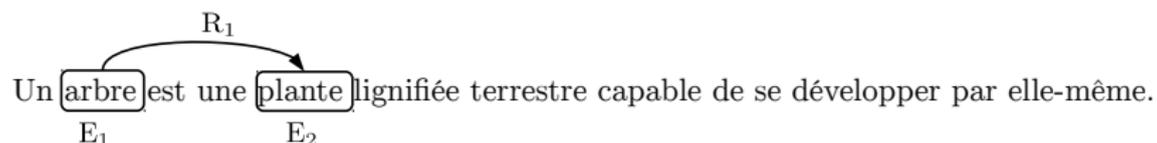
10 novembre 2014



### Extraction de Relations

- Tâche appartenant à l'extraction d'information (IE) et qui vise à extraire des **relations sémantiques** entre des **entités du monde**.
- Tâche cruciale pour la construction de ressources ou applications TAL :
  - e.g. : WordNet, DBpedia, etc.
  - e.g. : Recherche d'Information, Systèmes Question-Réponse, etc.

- **Exemple** : relation  $R_1$  d'hyponymie entre  $E_1$  et  $E_2$  :



- Une fois extraite, cette information est formalisée (FOL, DL, etc.) :
  - e.g. :  $\forall x(\text{arbre}(x) \implies \text{plante}(x))$
  - e.g. :  $\text{arbre} \sqsubseteq \text{plante}$

### Trois grandes familles de méthodes :

1. Approches symboliques,
2. Approches statistiques,
  - supervisées,
  - semi-supervisées,
  - non supervisées.
3. Approches mixtes.

### Facteurs influençant leur choix :

- évolution technologique (capacités de traitement, stockage, algorithmes),
- genre du corpus (journalistique, encyclopédique, etc.),
- nature des sources (structurées, non structurées),
- visée applicative (construction de ressources, etc.),
- utilisation de ressources (ressources distributionnelles, thésaurus, etc.).

# Introduction

## Extraction de relations

### Approche symbolique : patrons

Intuition : intégrer manuellement des connaissances linguistiques.

- Travaux : (hyperonymie) [Hearst, 1992], (méronymie) [Berland and Charniak, 1999], (multiples) [Condamines and Rebeyrolle, 1997] [Aussenac-Gilles and Jacques, 2008]
- Exemples de patrons :
  - Y tel que X
  - X et/ou autres Y
  - Y incluant X
  - X est une sorte de Y

*is-a(sérogroupe, test-de-présence-bactérienne)*

Des tests de présence bactérienne tels que le sérogroupe peuvent être utilisés dans certains cas.

# Introduction

## Extraction de relations

### Approches statistiques : supervisées

Intuition : intégrer de manière automatique les connaissances linguistiques au moyen de données annotées.

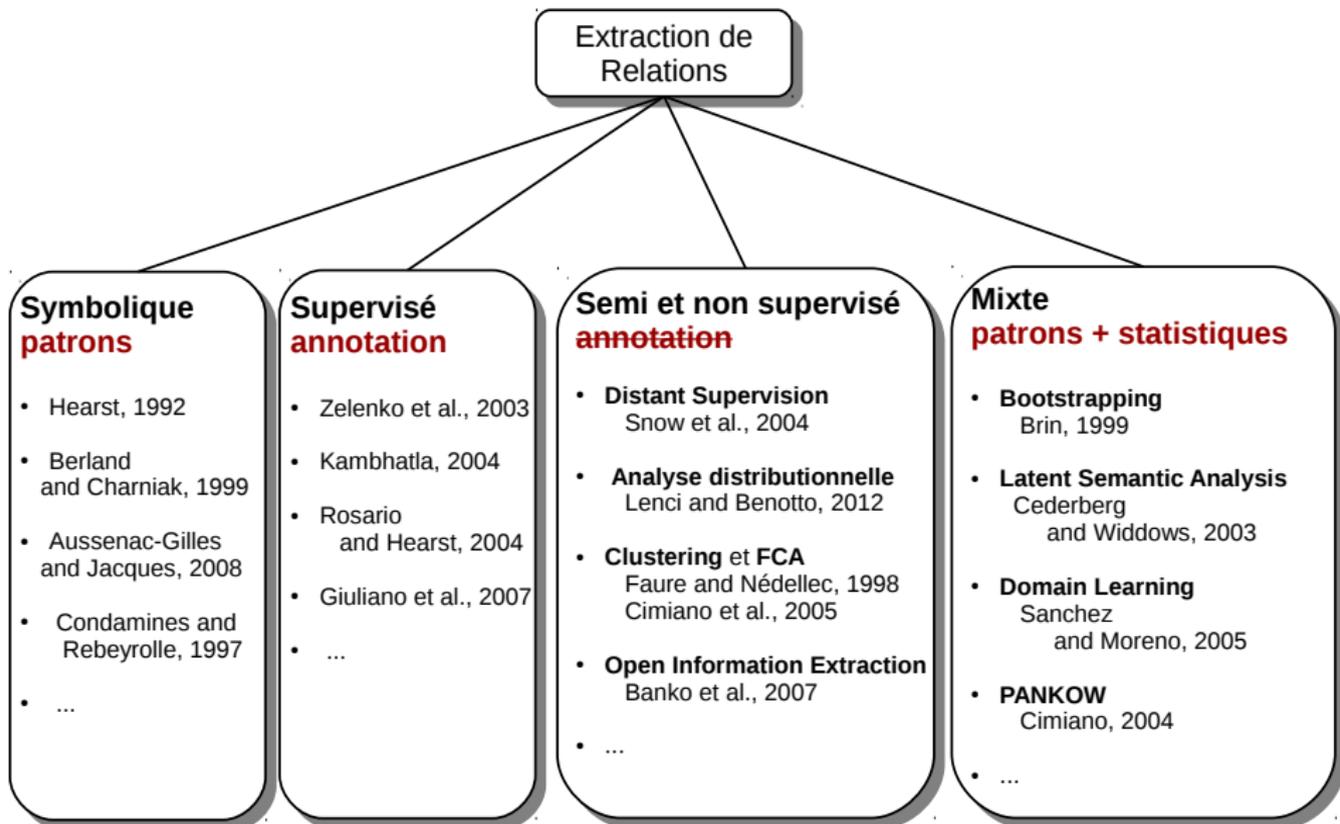
- Travaux : (svm) [Zelenko et al., 2003], (maxent) [Kambhatla, 2004], (neural network) [Rosario and Hearst, 2004]
- Deux sous-tâches :
  - Décider si 2 entités sont reliées,
  - Identifier la nature de la relation.
- Traits : contextes des entités, span entre les entités, types des entités, séquence des chunks, arbre de constituants et/ou de dépendances, etc.

*est-pdg-de(Guillaume Faury, Eurocopter)*

**Eurocopter** prendra un nouveau nom à partir de janvier 2014 dans le cadre d'une restructuration, a déclaré le PDG **Guillaume Faury**.

# Introduction

## Extraction de relations



# Introduction

## Dépasser le niveau de la phrase

### Dépasser le niveau de la phrase :

- Les approches actuelles travaillent au niveau de la phrase,
- Deux constats :
  1. Nombreuses relations exprimées au travers de la structure,
  2. Les outils actuels ne sont pas adaptés à certains objets textuels, e.g. : titres, listes à puces, tables, etc.
- ⇒ Nécessité d'adapter le traitement selon la nature de l'objet, son rôle et sa position dans le document.

### Vers un niveau du texte [Marcu, 2006]

« *From a natural language engineering perspective, the need for text-level processing systems is uncontroversial : **because sentence-level processing modules** (syntactic and semantic parsers, named-entity recognizers, language translators and generators, etc.) **operate at sentential level, they are not able to make text-level inferences** and/or produce outputs that are text-level coherent/consistent. »*

# Introduction

## Dépasser le niveau de la phrase

### La structure pour l'extraction de relations :

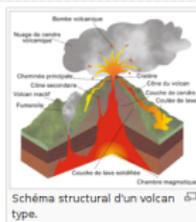
- La variation de la langue est plus grande quand la structure intervient,
- La structuration supplée/remplace l'aspect syntaxique [Bush, 2003],
- Existence d'une grande masse de données sur le Web.

#### Description

##### Structures et reliefs

Un volcan est formé de différentes structures que l'on retrouve en général chez chacun d'eux :

- une **chambre magmatique** alimentée par du magma venant du **manteau** et jouant le rôle de réservoir et de lieu de différenciation du magma. Lorsque celle-ci se vide à la suite d'une **éruption**, le volcan peut s'affaisser et donner naissance à une **caldeira**. Les chambres magmatiques se trouvent entre dix et cinquante kilomètres de profondeur dans la **lithosphère** <sup>5 (réf. insuffisante)</sup> ;
- une **cheminée volcanique** qui est le lieu de transit privilégié du magma de la chambre magmatique vers la surface ;
- un **cratère** ou une **caldeira sommitale** où débouche la cheminée volcanique ;
- une ou plusieurs **cheminées volcaniques secondaires** partant de la chambre magmatique ou de la cheminée volcanique principale et débouchant en général sur les **flancs** du volcan, parfois à sa base ; elles peuvent donner naissance à de petits cônes secondaires ;
- des **fissures latérales** qui sont des fractures longitudinales dans le flanc du volcan provoquées par son gonflement ou son dégonflement <sup>6 (réf. nécessaire)</sup> ; elles peuvent permettre l'émission de lave sous la forme d'une éruption fissurale.



##### Forme des volcans

La classification la plus courante dans les ouvrages de vulgarisation distingue **trois types de volcans** suivant le type de **lave** qu'ils émettent et le type d'**éruption** :

- en **volcan bouclier** lorsque son diamètre est très supérieur à sa hauteur en raison de la fluidité des laves qui peuvent parcourir des kilomètres avant de s'arrêter ; le **Mauna Kea**, l'**Erta Ale** ou le **Piton de la Fournaise** en sont des exemples <sup>7</sup> ;
- en **stratovolcan** lorsque son diamètre est plus équilibré par rapport à sa hauteur en raison de la plus grande viscosité des laves ; il s'agit des volcans aux éruptions explosives comme le **Vésuve**, le **mont Fuji**, le **Mont St Helens** <sup>8</sup> ;
- en **volcan fissural** formé par une ouverture linéaire dans la croûte terrestre ou océanique par laquelle s'échappe de la lave fluide ; les volcans des **dorsales** se présentent sous forme de fissure comme les **Lakagigar** ou le **Krafla** <sup>9</sup>.

Comme toute classification de phénomènes naturels, beaucoup de cas sont intermédiaires entre les types purs : l'**Etna** ressemble à un stratovolcan posé sur un volcan bouclier. **Hekla** est à la fois un stratovolcan et un volcan fissural. Les éruptions des volcans boucliers d'Hawaii démarrent souvent par l'ouverture d'une fissure. Dans *Volcanoes of the World*, Tom Simkin and Lee Siebert listent 26 types morphologiques. Sans aller aussi loin, on peut ajouter aux précédents **deux types** très différents <sup>9</sup> :

- les complexes de **caldeira rhyolitiques** comme la **caldeira de Yellowstone**, qui n'ont pas d'édifice volcanique,
- les **champs monogéniques**, qui présentent de multiples édifices comme des cônes de scories édifiés chacun en une seule fois.

et diviser les **volcans fissuraux** en :

- **trapps** et
- **dorsales océaniques**

# Introduction

## Dépasser le niveau de la phrase

### COGNITION LANGUES LANGAGE ERGONOMIE Équipe de Recherche en Syntaxe et Sémantique

#### Présentation

Les recherches de l'axe CARTEL portent sur le développement, l'adaptation et l'utilisation de ressources informatiques pour l'étude du langage. L'approche développée dans cet axe est essentiellement linguistique. Elle se fonde sur l'idée que l'observation de données massives fait émerger des questionnements linguistiques originaux et que l'exploitation de ressources riches et variées peut contribuer à leur traitement. Les travaux de l'axe visent à la fois :

- la construction de ressources langagières (corpus et lexiques) ;
- le développement d'outils d'ingénierie linguistique ;
- l'analyse linguistique outillée ;
- la modélisation du lexique.

Nos recherches abordent des domaines variés : la structuration du lexique sur les plans sémantiques et morphologiques, les aspects linguistiques de la recherche d'information, l'analyse syntaxique et discursive, la classification de documents, etc. Les productions (outils et données) sont rendues accessibles à la communauté, aussi librement que le permettent les contraintes de propriété, sur le serveur dédié REDAC (Ressources Développées À CLLE-ERSS) : [redac.univ-tlse2.fr](http://redac.univ-tlse2.fr)

Les trois principales thématiques de l'axe CARTEL sont :

- Sémantique distributionnelle
- Structuration du lexique
- Caractérisation et classification linguistique de corpus

#### Spécificités des travaux de l'axe

Au sein d'un environnement scientifique qui oriente le TAL vers un usage intensif de méthodes statistiques sur des données massives, nous privilégions le traitement linguistique. Notre approche donne une place centrale à l'observation fine des données manipulées avant et pendant un traitement automatisé, et au questionnement linguistique, même lorsque nous répondons à des besoins applicatifs. Ceci ne nous empêche pas de faire usage des méthodes par apprentissage (classification, fouille de données), en les articulant le plus souvent avec des modes d'observation outillés allant de la structure globale (visualisation, statistiques) aux phénomènes locaux (en facilitant le retour aux données initiales).

#### Liens avec l'enseignement

Les membres de l'axe CARTEL sont impliqués dans deux masters spécifiques :

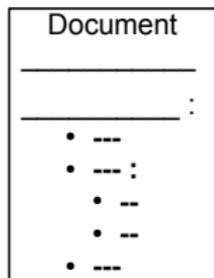
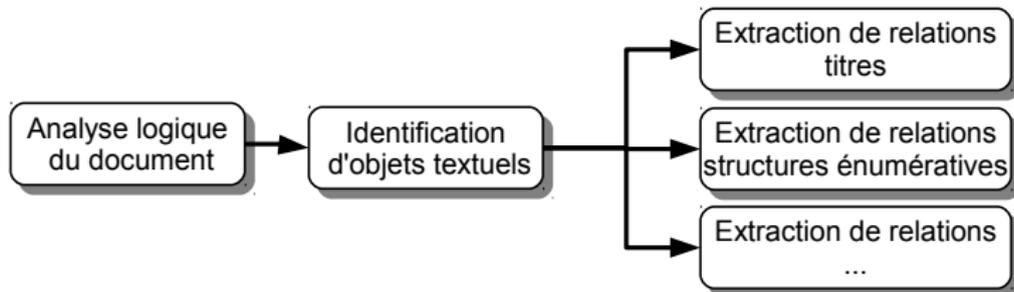
- Le master ECIL (Ergonomie Cognitive et Ingénierie Linguistique) à Toulouse
- Le master RELAI (Recherches Linguistiques et Applications Informatiques) à Bordeaux

# Introduction

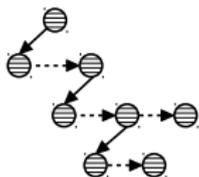
## Objectif et démarche

### Approche proposée :

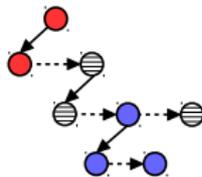
- 3 étapes :



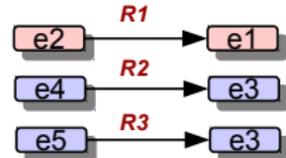
Arbre de dépendance



Identification d'objets textuels



Extraction de relations

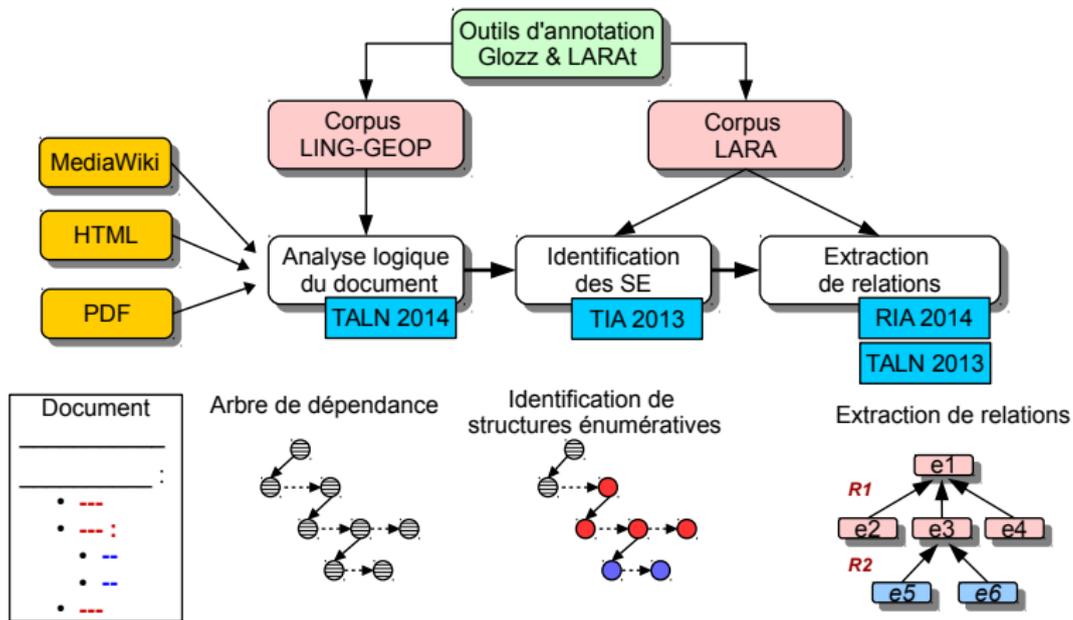


# Introduction

## Objectif et démarche

### Travail actuel :

- Utilisation des **structures énumératives** pour l'extraction de relations,
- **Apprentissage supervisé** et besoin de **données annotées**.



## 1 Document et l'Extraction de Relations

- Discours et Document
- Les niveaux de structure
- La structure logique profonde
- Lien avec l'extraction de relations

## 2 Analyse Logique du Document

## 3 Les Structures Énumératives

## 4 Conclusion et Perspectives

# Document et l'Extraction de Relations

## Un retour au document

### Discours et Document :

- Un « texte » est plus que la somme de ses phrases [Halliday and Hasan, 1976],
- Souvent les approches → « îlots textuels » [Nazarenko, 2005],
- Utilité d'une **cohérence** du document [Péry-Woodley and Scott, 2006],
- Le projet ANNODIS est un pas dans cette direction [Afantenos et al., 2012].

### Problématique du Workshop SCAR 2007

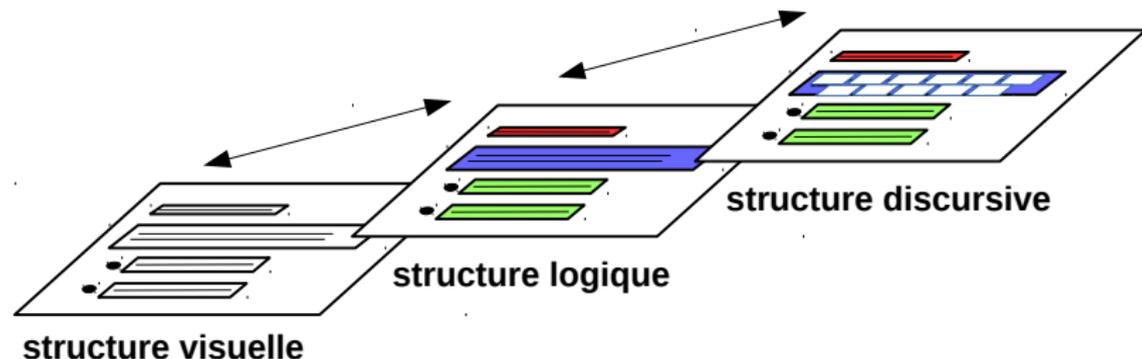
*Text (and language in general) has ABOUTNESS ; it has meaning, or semantic content. We as (computational) linguists are highly adept at dissecting text on a number of different levels : we can perform grammatical analysis of the words in the text, we can detect animacy and salience, we can do syntactic analysis and build parse trees of partial and whole sentences, and we can even identify and track topics throughout the text.*

*However, we are comparatively inept when it comes to identifying the semantic content, or meaning, of the text.*

# Document et l'Extraction de Relations

## Les niveaux de structure du document

Trois niveaux de structure du document :



### Segmentation en unités :

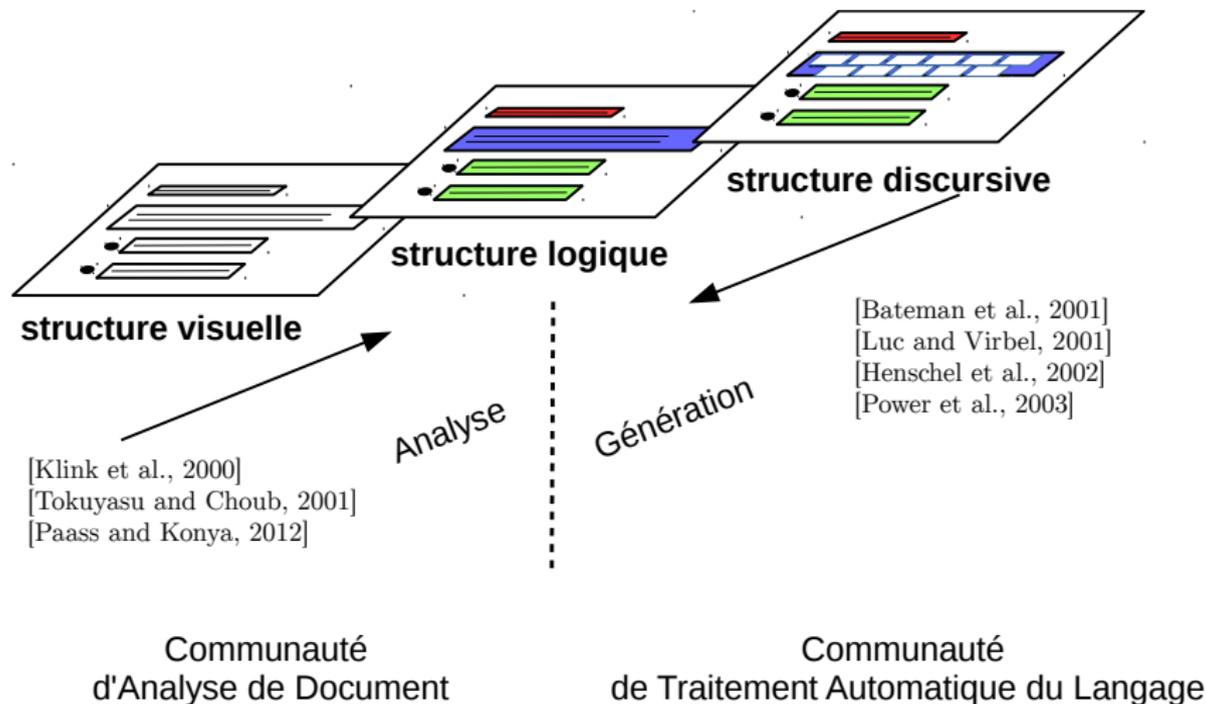
Consensus « flou » dans la littérature

- **Structure visuelle** : alinéa et blocs visuel, (//OCR)
- **Structure logique** : titre, paragraphe, item, etc. (//HTML,  $\text{\LaTeX}$ )
- **Structure discursive** : EDU et CDU.

# Document et l'Extraction de Relations

## Les niveaux de structure du document

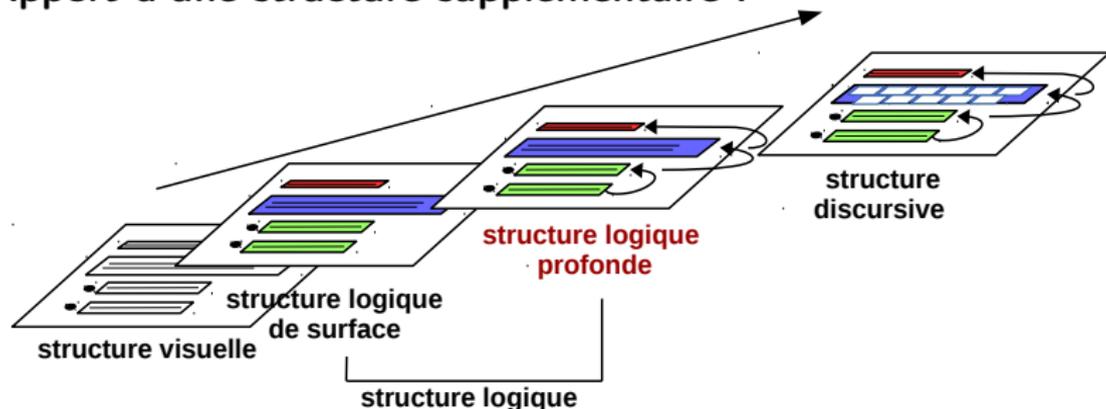
Littérature en structure du document :



# Document et l'Extraction de Relations

## Les niveaux de structure du document

Apport d'une structure supplémentaire :



Segmentation en unités :

- **Structure visuelle** : alinéa et blocs visuel, (//OCR)
- **Structure logique** :
  - **unités élémentaires** : titre, paragraphe, item, etc. (//HTML,  $\LaTeX$ )
  - **unités complexes** : section, structure énumérative, définition, etc.
- **Structure discursive** : EDU et CDU.

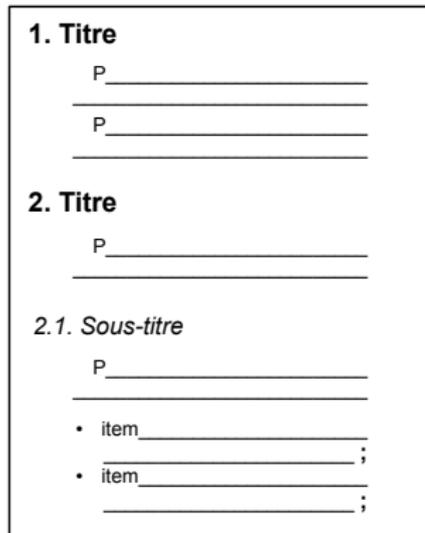
# Document et l'Extraction de Relations

## La structure logique profonde

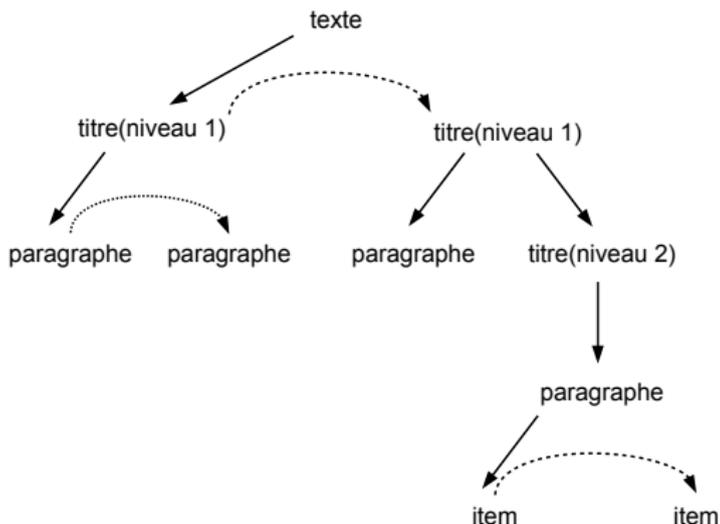
La structure logique profonde → un arbre de dépendance :

- Deux relations : subordination et coordination,
- Principe de dépendance : unités liées dans la **cohérence** du document,
- Composants : **nœuds** = unités élémentaires, **arcs** = dépendances typées.

document



structure logique profonde



# Document et l'Extraction de Relations

La structure logique profonde : constituants et dépendances

## Arbre de constituants et Arbre de dépendance :

### Arbre de constituants

- **phrase structure grammar** [Chomsky, 1957].
- Une unité d'un niveau est **constituée** d'unités de niveaux plus bas.

### Arbre de dépendance

- **dependency grammar** [Tesnière, 1959].
- Une unité est **dépendante** d'un gouverneur.

### Passage de l'un à l'autre

- Un arbre de constituants décrit implicitement des relations de dépendance,
- Une projection d'un sous-arbre de dépendance peut être vue en constituants,
- Cependant, le choix des constituants nécessite de déterminer a priori les labels des nœuds.
  - Ce qui complique la tâche dans l'analyse de données réelles.

### 2.1 Du mouvement à l'intention

L'intention associée au verbe *aller* concerne souvent les actes que l'auteur du mouvement se propose d'exécuter lorsqu'il arrive à la destination du mouvement:

- (3) Car incontinant le roy manda tous ses barons, cappitaines et cheffz de guerre, et sans aucun delay fit appareillier tout ce qui estoit de besoing pour *aller* en Espagne *commencer la guerre contre les barons du pays*. (Jehan de Paris 8, cité par Werner 1980: 131)

Il suffit que cette intention soit plus importante en contexte que la destination à laquelle elle est associée pour que seule l'intention soit exprimée. Detges (1999: 39) cite ainsi les exemples suivants, dans lesquels seule l'intention est encore explicitée,

- soit parce que le (co-)texte précise la destination du mouvement:

- (4) Nos *alomes* la messe *oir*; Tuit *alomes vers le mostier*. (*Roman de Renart* 12582, fin 12<sup>e</sup> – début 13<sup>e</sup> s.; cité par Littré 1961/62 et Detges 1999: 39)
- (5) Il meïsmes *ala trois serjans apeler* (*Li romans de Berte aus grans pies* XVII, fin 13<sup>e</sup> siècle, cité par Littré 1961/62 et Detges 1999: 39)

- soit parce que la destination peut être déduite à partir de nos connaissances encyclopédiques («toute action a lieu à un endroit particulier»):

- (6) Il est bien temps de deviser / Les personnaiges et nommer. / Je vous les *veulx* nommer à tous. / Je *voys* au Monde commencer. (*Moralité de Charité*, 1532-1550, passage cité par Gougenheim 1929/1971: 98 et Detges 1999: 39)

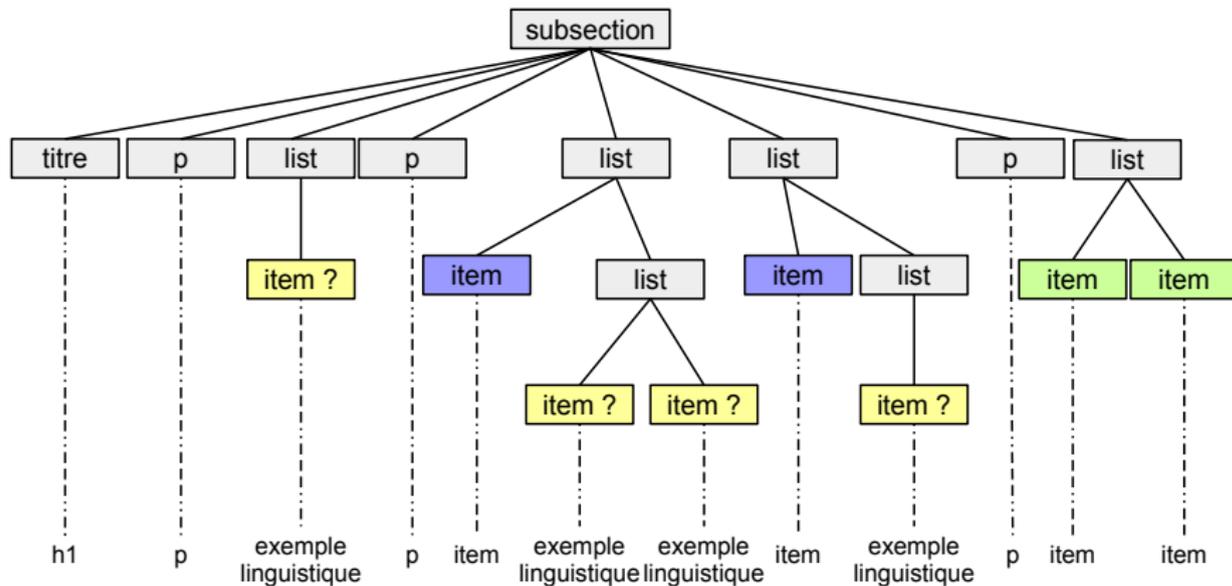
Bref, il n'est pas toujours nécessaire d'expliciter la destination parce que celle-ci peut être déduite sans problèmes du contexte. Partant, le verbe *aller* s'emploiera par la suite dans des contextes où il n'exprime plus l'idée d'un mouvement, mais où il signale seulement la présence d'une intention:

- **Le chat parti**, les souris dansent
- **Le chat étant parti**, les souris dansent,

# Document et l'Extraction de Relations

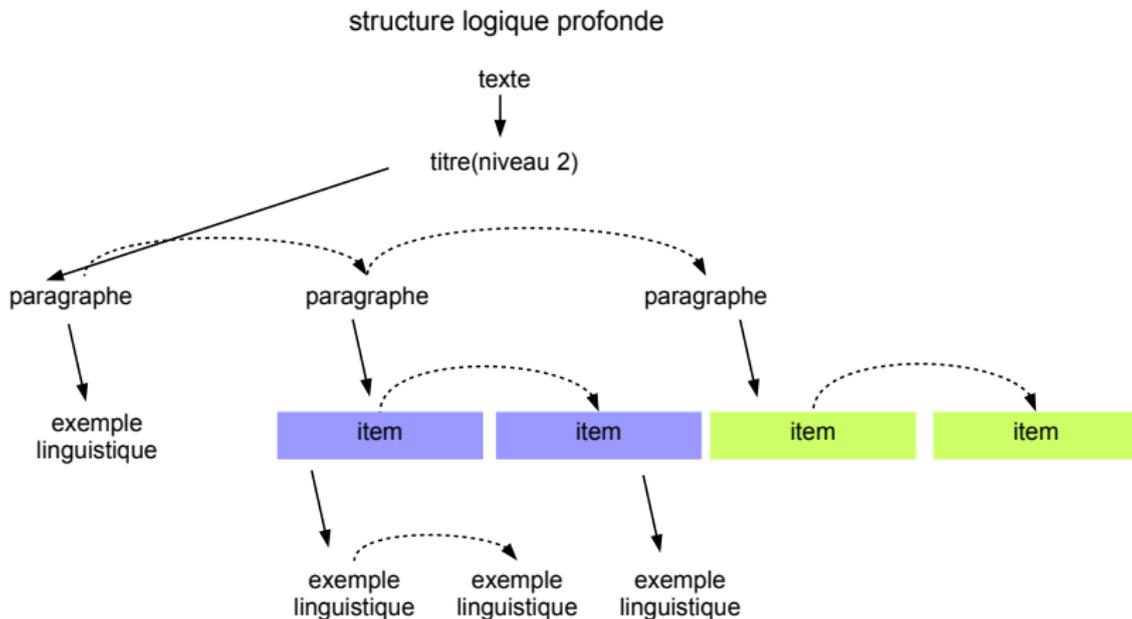
La structure logique profonde : constituants et dépendances

arbre de constituants



# Document et l'Extraction de Relations

La structure logique profonde : constituants et dépendances



# Document et l'Extraction de Relations

La structure logique profonde : deux relations

## Deux relations de dépendance :

*subordination* et *coordination*

- D'un point de vue discursif :
  - parallèle avec un modèle discursif simplifié,  
[Choi, 2000] [Hernandez and Grau, 2005],
  - adéquation avec les théories du discours,  
RST : [Mann and Thompson, 1988] :
    - *subordination* = *noyau-satellite*,
    - *coordination* = *multi-nucléaire*.
  - perspectives en parsing rhétorique.  
[Afantenos et al., 2010]
- D'un point de vue analyse du document :
  - construction s'apparente à du parsing LR.  
3 transitions : { *subordination*, *coordination*,  $\emptyset$  }
  - facilite l'identification d'objets textuels.  
support pour chevauchements, imbrications, etc.



### Les structures énumératives pour l'extraction de relations :

Deux cas rhétoriques [Luc, 2000] :

#### 1. Structures énumératives paradigmatiques :

- Énumérer : énoncer des éléments partageant un même champ conceptuel,
- Une relation identique unit tous les **items** à l'**amorçe**.

Un avion est constitué :

d'un groupe motopropulseur à hélice ou à réaction ;  
de commandes de vol permettant de transmettre ... ;  
de servitudes de bord, ensemble des circuits électriques, ... ;  
d'instruments de bord permettant ... .

#### 2. Structures énumératives syntagmatiques :

- les items ne sont pas coordonnés fonctionnellement,
- Il y a différentes relations en présence.

La course connut quatre épisodes :

Excalibur partit en tête, et bien que  
il soit quatrième dans le virage, et  
cinquième à l'entrée de la ligne droite,  
il l'emporta finalement d'une courte tête.

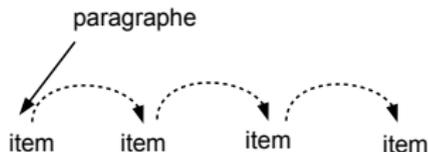
# Document et l'Extraction de Relations

## Isomorphisme entre structure logique profonde et Rhetorical Structure Theory

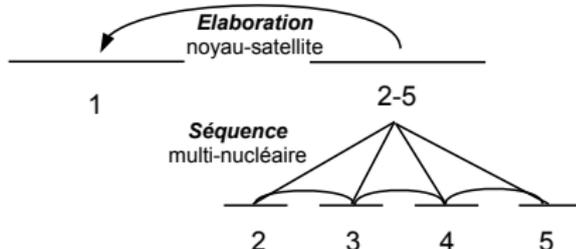
Texte :

Un avion est constitué :  
d'un groupe motopropulseur à hélice ou à réaction ;  
de commandes de vol permettant de transmettre ... ;  
de servitudes de bord, ensemble des circuits électriques, ... ;  
d'instruments de bord permettant ... .

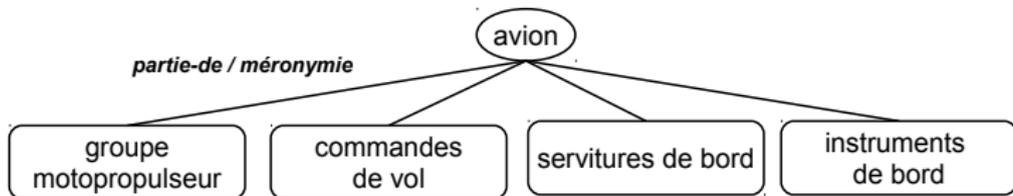
Structure logique profonde :



Rhetorical Structure Theory :



Représentation conceptuelle :



## 1 Document et l'Extraction de Relations

## 2 Analyse Logique du Document

- Démarche
- Corpus LING-GEOP
- Tâche 1
- Tâche 2

## 3 Les Structures Énumératives

## 4 Conclusion et Perspectives



# Analyse Logique du Document

## Démarche

### Deux tâches :

- (1) Identifier les **unités logiques élémentaires** dans le document,
- (2) Construire l'**arbre de dépendance** ordonnant ces unités.

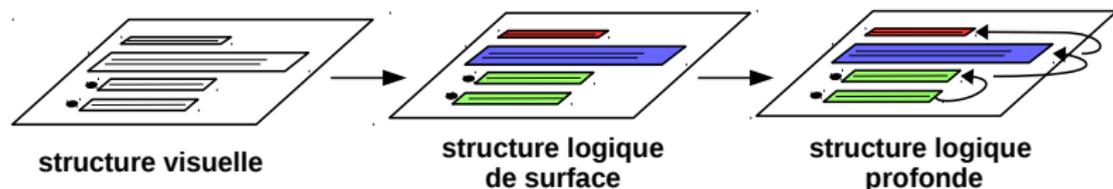
### Différents formats :

- MediaWiki : uniquement tâche 2 (déterministe),
- HTML : nettoyage et tâche 2 (déterministe),
- PDF : OCR, tâche 1 et tâche 2 (non déterministe).  
→ Nécessité de données annotées

[Pustejovsky and Stubbs, 2012]

*(...) it is not enough to simply provide a computer with a large amount of data and expect it to learn to speak—the data has to be prepared in such a way that the computer can more easily find patterns and inferences. This is usually done by adding **relevant metadata** to a dataset. Any metadata tag used to mark up elements of the dataset is called an **annotation** over the input.*

Nécessité d'une annotation visuelle et logique :



Le corpus LING-GEOP :

- Sous-ensemble de ANNODIS ME en PDF
- Annoté avec des structures multi-échelles
- Des genres différents :
  - **LING** : 25 articles issus du CMLF 2008
  - **GEOP** : 21 rapports/articles de l'IFRI
- Des natures différentes :
  - **LING** : Formaté visuellement et très structuré
  - **GEOP** : Pas de consensus visuel et peu structuré



### Segmentation en blocs visuels :

1. Utilisation de l'outil LA-PDFText [Ramakrishnan et al., 2012]  
PDF → blocs visuels
2. Correction manuelle des erreurs commises  
e.g : paragraphes coupés, inversions, etc.

### → Représentation des propriétés visuelles :

- Caractérisation dispositionnelle en pixels :  $(x_1, y_1)$   $(x_2, y_2)$
- Caractérisation typographique pour les mots : *police*, *style*, *contenu*.

```
<page x1="70" y1="71" x2="524" y2="806">  
  <chunk x1="70" y1="346" x2="524" y2="360">  
    <word x1="106" y1="346".. font="Arial" style="16pt;Bold">Le</word>  
    <word x1="135" y1="346".. font="Arial" style="16pt;It">sens</word>  
    ...  
  </chunk>  
</page>
```



# Analyse Logique du Document

Corpus LING-GEOP : (2) annotation de la structure logique de surface

## Annotation des labels logiques :

1. Réutilisation des labels présents dans ANNODIS  
Réalisée avec un algorithme de similarité textuelle [Myers, 1986]
2. Ajout manuel des labels non traités dans ANNODIS  
e.g : en-têtes, note de bas de page, etc.

## → Distributions pour LING et GEOP :

- Nombre équivalent de *paragraphes*
- Nombreux *items* dans LING
- Nombreux *autres* dans GEOP

	h	para.	item	cit.	en-tête	piéd p.	note p.	bibl	autres
LING	304	1241	380	123	45	16	394	1173	82
Moy.	12,1	49,6	<b>15,2</b>	<b>4,9</b>	1,8	0,6	15,7	<b>46,9</b>	3,2
GEOP	241	1189	72	1	171	257	122	398	195
Moy.	11,4	56,6	3,4	0,05	<b>8,1</b>	<b>12,2</b>	5,8	18,9	<b>9,2</b>

# Analyse Logique du Document

Corpus LING-GEOP : (3) annotation de la structure logique profonde

## Annotation de l'arbre de dépendance :

1. Ajout de subordination et coordination entre unités logiques  
Approche symbolique
2. Ajout des structures multi-échelles de ANNODIS ME  
Uniquement celles dont la granularité est supérieure au bloc visuel

## → Distributions pour LING et GEOP :

- Nombreuses subordinations et coordinations dans LING
- Prédominance générale des coordinations sur les subordinations

	subordination	coordination	Total
LING	714	2467	3181
Moy.	<b>28,56</b>	<b>98,68</b>	127,24
GEOP	391	1029	1420
Moy.	18,62	49	67,62
Couv.%	0,24	0,76	100%

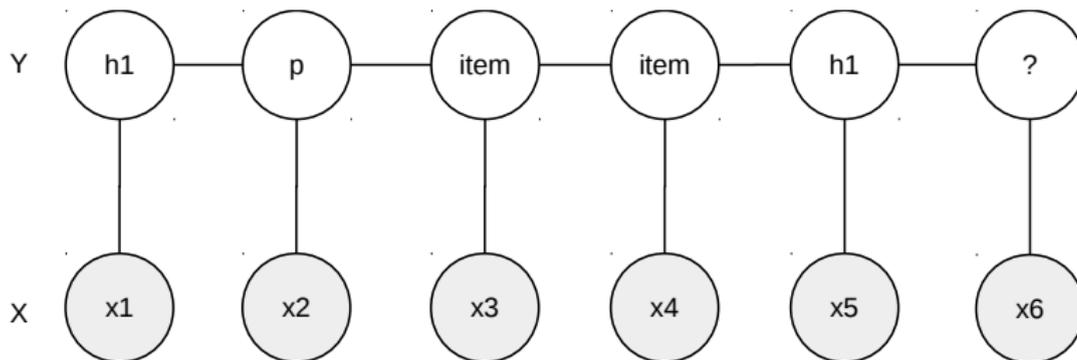
# Analyse Logique du Document

## Tâche 1 : identification des unités logiques élémentaires

### Tâche 1 :

Étiquetage des **blocs visuels** issus de LA-PDFText avec les labels des **unités logiques élémentaires** .

- Utilisation de la séquence des documents,  
Intuition : les documents présentent des régularités dans leur structure.
- Classifieur : Linear-Chain CRF



# Analyse Logique du Document

## Tâche 1 : traits

### Traits pour la Tâche 1 :

- Deux familles de traits : **locaux** et **de séquence**,
- Travail de discrétisation des valeurs.

Familles	Traits	Informations capturées
Traits locaux	<i>marges</i>	Indentation à droite ou à gauche, centrage des blocs, absence d'indentation, etc.
	<i>polices</i>	Présence d'emphases (gras ou italique), taille de la police, etc.
	<i>typographies</i>	Présence de puces, de tirets, de numérotation, d'un « ; » ou « , » en fin de bloc, etc.
	<i>positions</i>	Position verticale dans la page (haut, bas) et horizontale (droite, gauche)
	<i>ratios</i>	Ratios de la surface sur la taille de la police, de longueur sur la largeur, etc.
Traits de séquence	<i>bigrammes</i>	Considère le label <i>y</i> attribué à l'unité qui précède dans la séquence du document
	<i>debuts/fins</i>	Présence du bloc en début ou en fin de document
	<i>contrastes</i>	Rupture avec le bloc qui précède (taille, type de police, indentation, etc.)

# Analyse Logique du Document

## Tâche 1 : résultats

### Résultats pour la Tâche 1 (exactitude) :

Approches	LING	GEOP	LING_GEOP
Traits locaux	78,37%	79,97%	73,63%
+ Traits de séquence	<b>87,18%</b>	<b>82,39%</b>	<b>80,46%</b>
Baseline naïve	32,33%	44,51%	37,33%

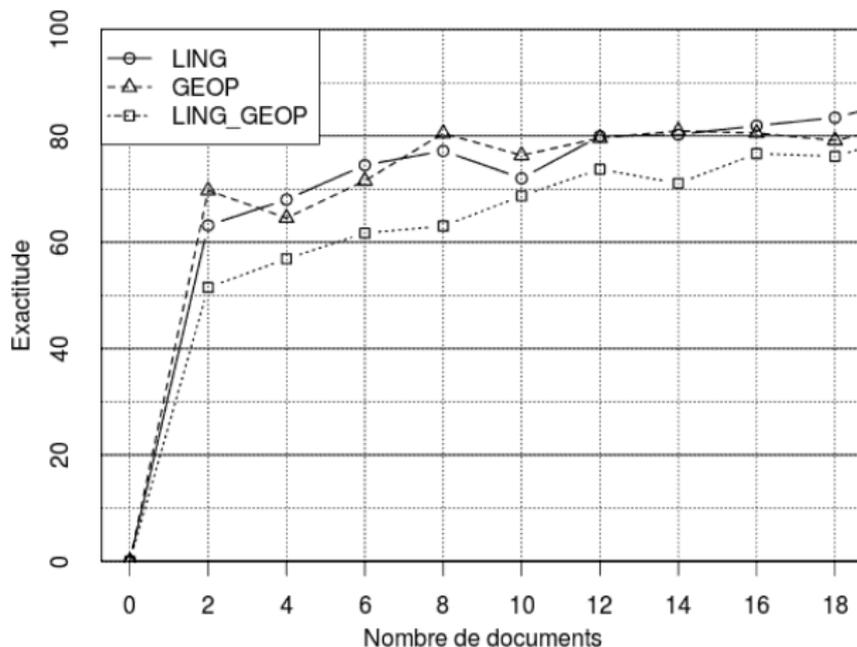
- **Baseline** :
  - classification en paragraphes (classe majoritaire)
- **Traits locaux** :
  - LING présente des objets ambigus visuellement (exemples linguistiques)
- **Traits de séquence** :
  - La séquence de labels permet de gérer la variation locale des unités
  - GEOP présente une structure simple et une classe *autres* forte
    - items (F-score) : 67,59 LING vs. 26,47 GEOP
    - titres niveau 2 (F-score) : 94,45 LING vs. 53,17 GEOP

# Analyse Logique du Document

## Tâche 1 : résultats

### Courbes d'apprentissage :

- Expériences modifiant la taille du set d'apprentissage
- LING\_GEOP : diminution des différences de distributions



# Analyse Logique du Document

## Tâche 2 : construction de l'arbre de dépendance

### Tâche 2 :

Relier les **unités logiques élémentaires** par des coordinations et subordinations pour former un **arbre de dépendance**.

- Parsing *shift-reduce* [Hernandez and Grau, 2005]
- Classifieur : Maximum d'Entropie

---

#### algorithm 1 : shift-reduce parsing

---

```
1: push root on  $\sigma$ 
2: while  $\beta$  and  $\sigma$  are not empty :
3:   if  $\text{arc}[\sigma_0, \beta_0] == \text{subordination}$  :                               /*reduce*/
4:      $a \leftarrow \beta_0$  and pop  $\beta_0$ 
5:     push  $a$  on  $\sigma$ 
6:   else if  $\text{arc}[\sigma_0, \beta_0] == \text{coordination}$  :                       /*reduce*/
7:      $a \leftarrow \beta_0$ 
8:     pop  $\sigma_0$  and  $\beta_0$ 
9:     push  $a$  on  $\sigma$ 
10:  else                                                                    /*shift*/
11:    pop  $\sigma_0$ 
```

# Analyse Logique du Document

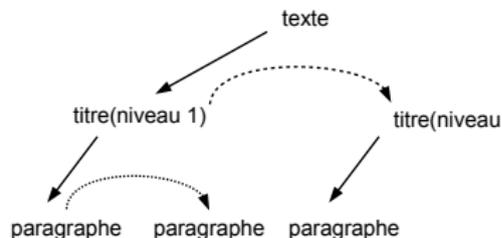
## Tâche 2 : exemple de parsing

### Exemple de parsing :

Pour un document :

**h1,p,p,h1,p.**

Nous voulons obtenir l'arbre  $\Rightarrow$



	transition	pile $\sigma$	liste $\beta$	dépendances
0		root	h1, p, p, h1, p	
1	<i>reduce</i>	root, h1	p, p, h1, p	sub(h1, root)
2	<i>reduce</i>	root, h1, p	p, h1, p	sub(p, h1)
3	<i>reduce</i>	root, h1, p	h1, p	coord(p, p)
4	<i>shift</i>	root, h1	h1, p	
5	<i>reduce</i>	root, h1	p	coord(h1, h1)
6	<i>reduce</i>	root, h1, p	$\emptyset$	sub(p, h1)
7	<i>shift</i>	root, h1	$\emptyset$	
8	<i>shift</i>	root	$\emptyset$	
9	<i>shift</i>	$\emptyset$	$\emptyset$	

# Analyse Logique du Document

## Tâche 2 : traits

### Tâche 2 :

- 4 types de traits utilisés : **visuels**, **lexicaux**, **labels** et **parallélismes**,
- Les traits **visuels** sont identiques à la Tâche 1,
- Hypothèse que l'on dispose des labels de la Tâche 1.

Traits	Informations capturées
<i>visuels</i>	Présence d'indentation, de tirets, de puces, de « : », etc.
<i>lexicaux</i>	Présence de marqueurs d'intégration linéaire (e.g : <i>D'une part, D'autre part</i> , etc.)
<i>labels</i>	Paires de labels (e.g : titre-paragraphe, item-item, paragraphe-item, etc.) et égalité de labels
<i>parallélismes</i>	Paragraphe entre deux items visuellement identiques, deux items mais différents visuellement, etc.



### Résultats pour la Tâche 2 (exactitude) :

Approches	LING	GEOP	LING_GEOP
Traits	96,41%	<b>98,45%</b>	<b>97,23%</b>
Grammaire	<b>96,54%</b>	98,30%	97,08%
Baseline naïve	40,21%	41,03%	39,79%

- **Baseline** :
  - classification aléatoire (subordination et coordination)
- **Traits** :
  - LING présente une structuration complexe
  - subordination (F-score) : 91,99 LING vs 97,15 GEOP
  - coordination (F-score) : 97,69 LING vs 98,93 GEOP
- **Grammaire** :
  - Les relations entre unités suivent majoritairement la grammaire
  - Cette asymétrie induit un apprentissage de la grammaire
  - Les traits considérés comme discriminants sont ignorés

### Deux stratégies pour l'évaluation des Traits :

Évaluation de l'approche par traits sur deux sous-ensembles :

- Sous-ensemble des erreurs de la grammaire
- Sous-ensemble des unités suivant la grammaire

Stratégies	LING	GEOP	LING_GEOP
Traits sur erreurs grammaire	14,54% (16/110)	16,66% (4/24)	14,17% (19/134)
Traits hors erreurs grammaire	99,34% (3051/3071)	99,85% (1394/1396)	99,73% (4455/4467)

- **Traits sur erreurs grammaire :**
  - (Très) léger gain qui reste stable sur les corpus
- **Traits hors erreurs grammaire :**
  - 20 erreurs pour LING (sur 3071)
  - 2 erreurs pour GEOP (sur 1396)
  - 12 erreurs pour LING\_GEOP (sur 4467)



### Discussion :

- Utilisation conjointe est sensible au bruit,
- Tâche 1 : dépendance au corpus pour le choix des labels,
- Tâche 2 : apport des traits lexicaux non significatif :
  - Grain limité aux blocs visuels,
  - Présence limitée dans ANNODIS,  
Aspect visuel supplée souvent l'aspect lexical
  - Couverture et variabilité.

### Perspectives :

- Non-supervisé pour la Tâche 1,
- Utilisation de traits syntaxiques pour la Tâche 2,
- Un travail sur le parallélisme pour la Tâche 2.



- 1 Document et l'Extraction de Relations
- 2 Analyse Logique du Document
- 3 Les Structures Énumératives**
  - Démarche
  - Corpus LARAt
  - Caractérisation de la relation
  - Identification des entités
- 4 Conclusion et Perspectives



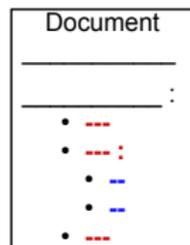
# Les Structures Énumératives

## Démarche

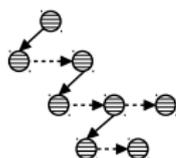
### Deux tâches pour l'extraction de relations :

- (1) Identifier la **nature de la relation**,  
e.g : *est-un*, *partie-de*, etc.
- (2) Identifier les **entités**,  
e.g : « avion », « ailes », « moteur », etc.

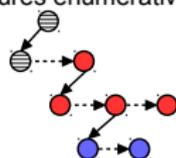
→ besoin de données annotées.



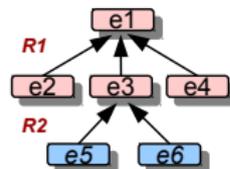
Arbre de dépendance



Identification de structures énumératives



Extraction de relations



### Corpus LARA :

- Débuté en juin 2013,
- Trois besoins :
  1. Des données pour l'apprentissage et l'évaluation,
  2. Une meilleur caractérisation des SE,
  3. Éprouver une nouvelle typologie.  
précédentes : [Virbel, 1999], [Luc, 2001], [Ho-Dac et al., 2010].
- Cadre :
  - 190 documents,
  - 3 annotateurs,
  - 1 guide d'annotation,
  - 1 outil d'annotation (LARAt).
- Base du corpus : Wikipedia  
Pages des concepts de l'ontologie OntoTopo.

# Les Structures Énumératives

Corpus LARA : annotation

## Démarche d'annotation :

1. Localisation de la SE dans le document,
2. Catégorisation selon axe rhétorique, intentionnel, sémantique,
3. Localisation interne de l'amorce, des items, des concepts.

## Axe sémantique :

- **Ontologique** : lien entre entités du monde,  
e.g. : « avion » et « véhicule ».
- **Lexical** : lien entre signifiants du langage,  
e.g. : « base de données » et « base chimique ».
- **Autre** : SE narratives, etc. → relations rhétoriques.

visée ontologique	lexical	autre sémantique
isA	homonymie	sémantiqueAutre
partOf	synonymie	
instanceOf	multilingue	
ontologiqueAutre	lexicalAutre	



# Les Structures Énumératives

## LARAt : un outil d'annotation pour les SE

File Credits

Text

risques des volcans.  
Le terme « volcan » tire son origine de **Vulcano**, une des **Sept collines** nommée en l'honneur de **Vulcan**, le **dieu romain** du feu dont l'équivalent dans le **panthéon grec** est **Héphaïstos**.

### Description

#### Structures et reliefs

Un volcan est formé de différentes structures que l'on retrouve en général chez chacun d'eux :

- une **chambre magmatique** alimentée par du **magma** venant du **mantau** et jouant le rôle de réservoir et de lieu de différenciation du magma. Lorsque celle-ci se vide à la suite d'une **éruption**, le volcan peut s'affaisser et donner naissance à une **caldera**. Les chambres magmatiques se trouvent entre dix et cinquante kilomètres de profondeur dans la **lithosphère** ;
- une **cheminée volcanique** qui est le lieu de transit privilégié du magma de la chambre magmatique vers la surface ;
- un **cratère** ou une **caldera** sommitale où débouche la cheminée volcanique ;
- une ou plusieurs **cheminées volcaniques secondaires** partant de la chambre magmatique ou de la cheminée volcanique principale et débouchant en général sur les flancs du volcan, parfois à sa base ; elles peuvent donner naissance à de petits cônes secondaires ;
- des **failures latérales** qui sont des fractures longitudinales dans le flanc du volcan provoquées par son gonflement ou son dégonflement ; elles peuvent permettre l'émission de lave sous la forme d'une éruption fissurale.

#### Forme des volcans

La classification la plus courante dans les ouvrages de vulgarisation distingue trois types de volcans suivant le type de **lave** qu'ils émettent et le type d'**éruption** :

- un **volcan bouclier** lorsque son diamètre est très supérieur à sa hauteur en raison de la fluidité des laves qui peuvent parcourir des kilomètres avant de s'arrêter : le **Mauna Kea**, l'**Éta Aie** ou le **Piton de la Fournaise** en sont des exemples ;
- un **stratovolcan** lorsque son diamètre est plus équilibré par rapport à sa hauteur en raison de la plus grande viscosité des laves ; il s'agit des volcans aux éruptions explosives comme le **Vésuve**, le **mont Fuji**, le **Merapi** ou le **mont Saint-Helens** ;
- un **volcan fissural** formé par une ouverture linéaire dans la croûte terrestre ou océanique par laquelle s'échappe de la lave fluide ; les volcans des **dorsales** se présentent sous forme de fissure comme le **Laki** ou le **Krafla**.

Comme toute classification de phénomènes naturels, beaucoup de cas sont intermédiaires entre les types purs : l'**Etna** ressemble à un stratovolcan posé sur un volcan bouclier, **Hekla** est à la fois un stratovolcan et un volcan fissural, les éruption des volcans boucliers d'Hawaï démarrent souvent par l'ouverture d'une fissure. Dans *Volcanoes of the World*, Tom Simkin and Lee Siebert listent 26 types morphologiques. Sans aller aussi loin, on peut ajouter aux précédents deux types

### Structures et reliefs

Un volcan est formé de différentes structures que l'on retrouve en général chez chacun d'eux :

- une **chambre magmatique** alimentée par du **magma** venant du **mantau** et jouant le rôle de réservoir et de lieu de différenciation du magma. Lorsque celle-ci se vide à la suite d'une **éruption**, le volcan peut s'affaisser et donner naissance à une **caldera**. Les chambres magmatiques se trouvent entre dix et cinquante kilomètres de profondeur dans la **lithosphère** ;
- une **cheminée volcanique** qui est le lieu de transit privilégié du magma de la chambre magmatique vers la surface ;
- un **cratère** ou une **caldera** sommitale où débouche la cheminée volcanique ;
- une ou plusieurs **cheminées volcaniques secondaires** partant de la chambre magmatique ou de la cheminée volcanique principale et débouchant en général sur les flancs du volcan, parfois à sa base ; elles peuvent donner naissance à de petits cônes secondaires ;
- des **failures latérales** qui sont des fractures longitudinales dans le flanc du volcan provoquées par son gonflement ou son dégonflement ; elles peuvent permettre l'émission de lave sous la forme d'une éruption fissurale.

Selection

Internal components

Status

Amorce : 1  
Items : 5  
Visual axis :  
Rhető axis :  
Inten axis :  
Sem axis :

Internal Units

Axe visuel

verticale  
 horizontale  
 hypertextuelle  
 navigationnelle

Axe rhétorique

Paradigmatique  
 Syntagmatique  
 Hybride  
 Bivalente

Axe intentionnel

descriptive  
 explicative  
 procédurale  
 autre intent.

Axe sémantique

A visée ontologique

isA  
 partOf  
 instanceOf  
 ontologiqueAutre

Metalinguistique

hyperonymie  
 méronymie  
 homonymie  
 synonymie  
 multilingue  
 lexicalAutre

Autre sémantique

sémantiqueAutre

Contexte

contextuelle  
 non contextuelle

Commentaire

Validation

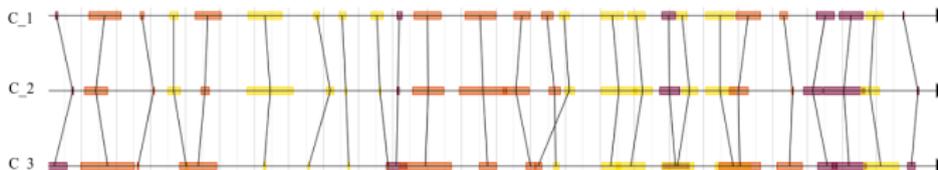
Control

ID SE : 1  
Nb. SE : 10  
Opérateur : AL

previo... P...

### Alignement et accord sur des unités non-prédéfinies :

- À partir de quel écart est-on prêt à dire que 2 unités ayant une **position** légèrement différente désignent bien le même phénomène linguistique ?
- L'**accord catégoriel** nécessite que les unités soient alignées **positionnellement**.



### Démarche suivie :

1. Génération d'alignements unitaires ( $\sum_{i=1}^m (\prod_i^n card_i)$ )
2. Mesure de l'écart absolu en caractères,
3. Solution approchée [Mathet and Widlöcher, 2011],
4. Nettoyage manuel des alignements unitaires,
5. Accord positionnel et Accord catégoriel.



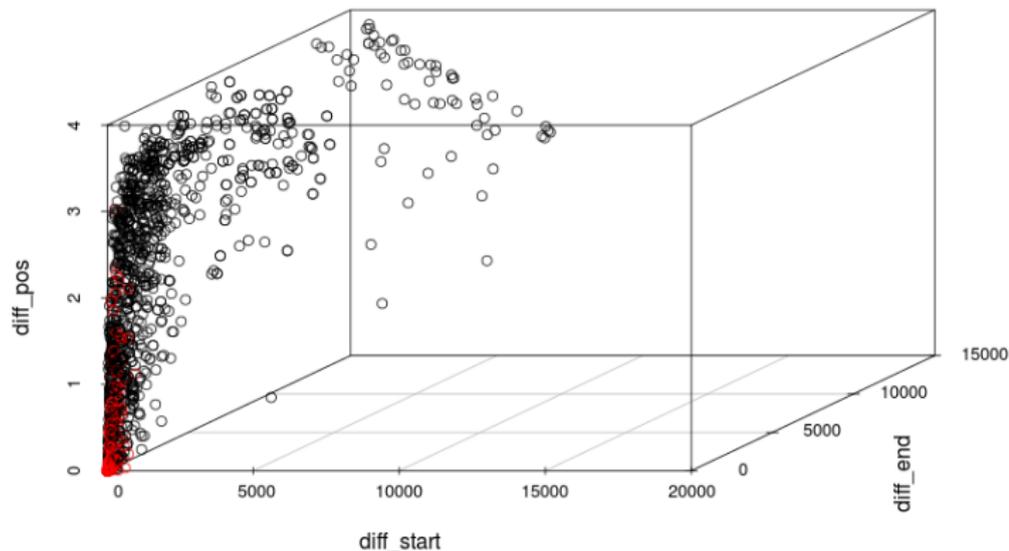


# Les Structures Énumératives

Corpus LARA : alignement

## Alignements unitaires après nettoyage :

- Rouge = alignements unitaires corrects,
- Noir = alignements unitaires incorrects.



# Les Structures Énumératives

Corpus LARA : accord

## Mesures de l'accord :

- Position des SE : F-score 83,21
- Position des entités (amorce) : F-score 77,27
- Position des entités (items) : F-score 81,61
- Axe Sémantique (verticales) :  
Kappa de Fleiss [Fleiss, 1971]

Classes	$\kappa$	z-score	observations	couverture
<i>isA</i>	<b>0.45</b>	18.27	268	36.0%
<i>instanceOf</i>	<b>0.43</b>	17.40	196	26.3%
<i>partOf</i>	<b>0.48</b>	19.53	39	5.2%
<i>otherOntological</i>	0.28	11.39	42	5.7%
<i>lexical</i>	0.74	30.40	149	20.0%
<i>other</i>	0.23	09.50	51	6.8%
Corpus	0.49	36.20	745	100%

# Les Structures Énumératives

## Identification de la relation

### Tâche 1 :

Identification de la **nature de la relation** entre l'amorce et l'énumération.

- Utilisation de traits de surface ("shallow features"),  
Parsing syntaxique pas adapté et passage à l'échelle plus aisé.
- Classifieurs : MaxEnt (log-linéaire) et SVM (noyau gaussien).

Pour un transformateur triphasé, il existe 3 types de couplage d'enroulement :

[le] **couplage étoile**, défini par la lettre Y ;

[le] **couplage triangle**, défini par la lettre D ou  $\Delta$  ;

[le] **couplage zig-zag**, défini par la lettre Z.

→ *hyperonymie*

Le déroulement

[Mélanger] la farine, le sucre, le sucre vanillé, les œufs, l'huile et le lait.

[Verser] la pâte dans la poêle, cuire un côté et retourner la crêpe.

→ *non-hyperonymie*

# Les Structures Énumératives

## Identification de la relation

### Approches précédentes :

1. Approche symbolique [Kamel and Rothenburger, 2011].
2. Tri-grammes de PosTag et de lemmes :  
Pour chaque tri-grammes de tokens,  $2^3$  traits :  
Le-DET chat-NC noir-ADJ  
DET chat-NC noir-ADJ  
...  
DET NC ADJ
3. Classification multi-classes :  
Le même modèle pour tous les types de relation.
4. Système de vote sur les classification des paires amorce-item.

### Approche proposée

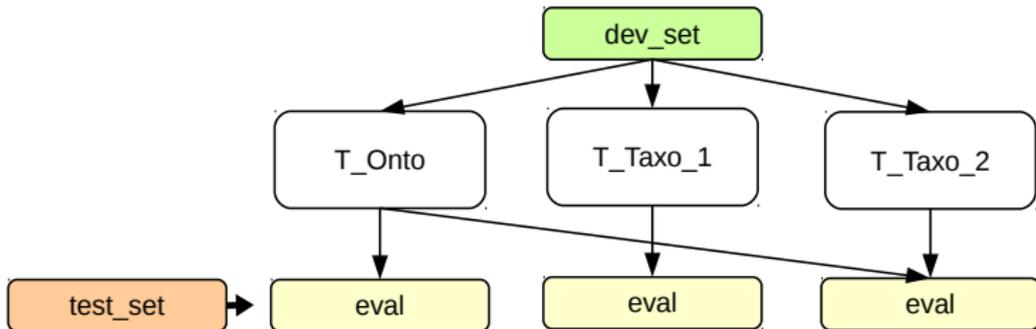
- Travail d'analyse et de sélection des traits,
- Un modèle par type de relation,
- Utilisation conjointe de modèles.

# Les Structures Énumératives

## Identification de la relation

### Trois modèles :

1. **T\_Onto** : ontologique vs. non-ontologique.
2. **T\_Taxo\_1** : hyperonymie vs. non-hyperonymie
3. **T\_Taxo\_2** : hyperonymie vs. non-hyperonymie + sorties de T\_Onto.



### Traits :

- **Amorce** : PosTag, position des verbes, marques du pluriel, ponctuation, nombres de phrases, de tokens, marqueurs de relation, trou syntaxique.
- **Item** : PosTag, position des verbes, nombres de sentences, de tokens.

# Les Structures Énumératives

Identification de la relation : traits

## Analyse des traits pour la classe ontologique :

Ordonnement selon la valeur absolue corrélation de Pearson entre le vecteur **f** de traits et le vecteur **y** de classes.

	Traits	Segment	<i>r</i> corr.
1	nombreTokens=1	Amorce	-0.236
2	postag=V	Item	-0.219
3	NC pluriel	Amorce	0.210
4	postag=PREP	Item	0.210
5	postag=V	Amorce	0.195
6	postag=NPP	Amorce	0.176
7	marqueurs de composition	Amorce	0.151
8	SE avec trou syntaxique	Amorce	0.141
9	marqueurs linguistiques	Amorce	-0.126
10	nombreTokens=3	Item	0.099



# Les Structures Énumératives

Identification de la relation : traits

## Analyse des traits pour la classe hyperonymie :

Ordonnement selon la valeur absolue corrélation de Pearson entre le vecteur **f** de traits et le vecteur **y** de classes.

	Traits	Segment	<i>r</i> corr.
1	postag=V	Item	-0.259
2	commence par DET	Item	0.235
3	nombreTokens=5	Item	0.147
4	postag=NPP	Item	0.132
5	commence par N	Item	0.128
6	NC pluriel	Amorce	0.120
7	postag=NPP	Amorce	0.120
8	commence par VINF	Item	-0.113
9	marqueurs linguistiques	Amorce	-0.112
10	nombreTokens=3	Item	0.107



# Les Structures Énumératives

Identification de la relation : résultats pour classe ontologique

## Résultats pour la classe ontologique :

test_set	Précision	Rappel	F1	Exactitude
MaxEnt	80.65	93.46	86.58	78.77
SVM	79.84	96.26	<b>87.29</b>	<b>79.45</b>
baseline	73.28	100.0	84.58	73.28

- **Baseline** :
  - classification par défaut dans la classe ontologique,
  - Précision élevée due à la distribution du corpus.
- **MaxEnt** et **SVM** :
  - Gain face à la baseline,
  - SVM est significativement meilleur ( $p < 0.03$ ).



# Les Structures Énumératives

Identification de la relation : résultats pour classe hyperonymie

## Résultats pour la classe hyperonymie :

test_set	Précision	Rappel	F1	Exactitude
MaxEnt	70.59	78.26	74.23	65.75
SVM	71.05	88.04	78.64	69.86
MaxEnt+	78.00	84.78	81.25	<b>75.34</b>
SVM+	74.77	90.22	<b>81.77</b>	74.66
baseline	63.01	100.0	77.31	63.01

test_set	$\Delta$ F1	$\Delta$ Exactitude	p-valeur
MaxEnt+ vs MaxEnt	7.02	9.59	< 0.01
MaxEnt+ vs baseline	3.94	12.33	< 0.01
SVM+ vs SVM	3.13	4.79	< 0.02
SVM+ vs baseline	4.46	11.64	< 0.01

- **MaxEnt+** et **SVM+** :
  - Significativement meilleurs que la baseline,
  - Différence significative avec **MaxEnt** et **SVM**,
  - Inversion des tendances de précision et rappel,
  - Classifieur non-linéaire → risque de mauvaise généralisation.



# Les Structures Énumératives

## Identification des entités

### Tâche 2 :

Identification des **paires d'entités** entre l'amorce et l'énumération.

- Utilisation d'extracteurs terminologiques,
  - YaTeA [Aubin and Hamon, 2006]
  - ACABIT [Daille, 1996]
- Classifieur : MaxEnt (probabiliste)

Un **volcan** est formé de **différentes structures** que l'on retrouve en général chez chacun d'eux :

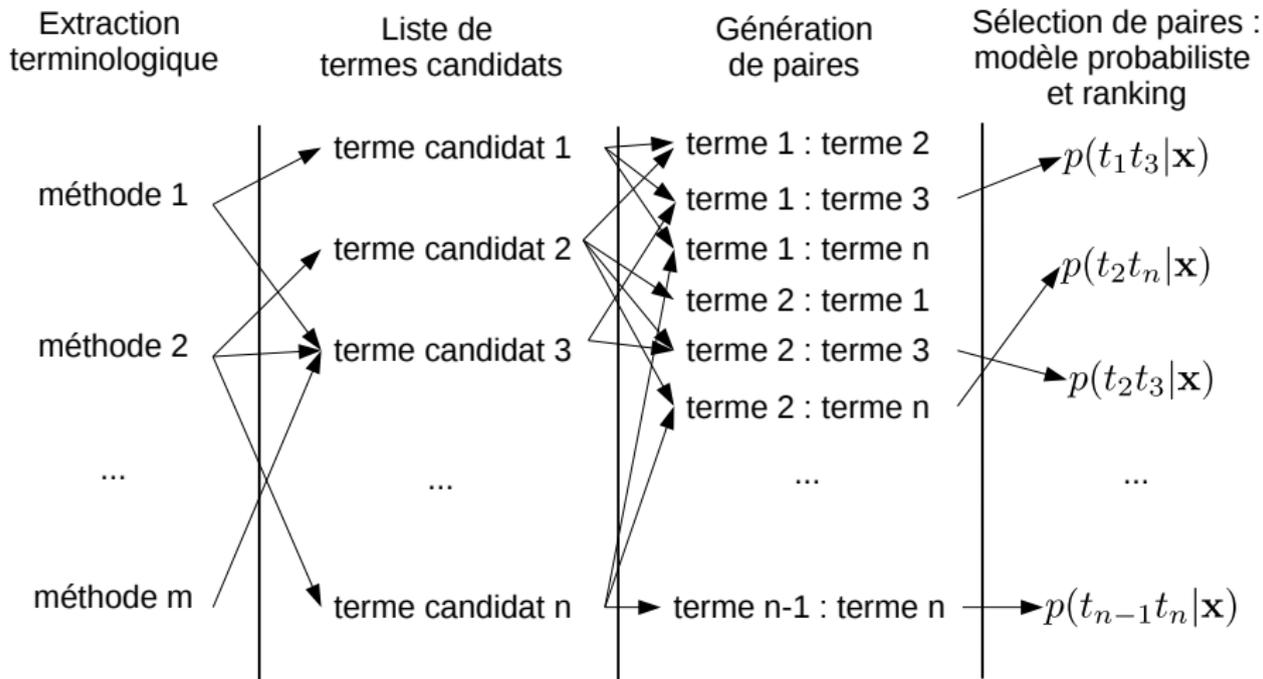
- une **chambre** magmatique alimentée par du magma venant du **mantau** et jouant le **role de reservoir et de lieu de differenciation** du magma. Lorsque celle-ci se **vide à la suite** d'une **eruption** le **volcan** peut s'effaisser et donner **naissance** à une **caldeira**. Les **chambres magmatiques** se trouvent entre **dix et cinquante** **kilometres de profondeur** dans la **litosphere**;
- une **cheminée** volcanique qui est le **lieu de transit privilégié** du **magma de la chambre magmatique** vers la **surface**;
- un **cratere** ou une **caldeira** **commitale** où débouche la **cheminée** volcanique;
- une ou plusieurs **cheminées volcaniques secondaires** partant de la **chambre magmatique** ou de la **cheminée volcanique principale** et débouchant en général sur les **flancs du volcan** parfois à sa **base**; elles peuvent donner **naissance** à de petits **cônes secondaires**;
- des **fissures latérales** qui sont des **fractures longitudinales** dans le **flanc du volcan** provoquées par son **gonflement** ou son **dégonflement**; elles peuvent permettre **l'émission de lave** sous la **forme** d'une **eruption fissurale**.



# Les Structures Énumératives

## Identification des entités : démarche

### Démarche :



# Les Structures Énumératives

Identification des entités : travail en cours

## Travail en cours :

- Traits : position des termes candidats, longueur des termes candidats, séquence de PosTag, inclusion lexicale, marqueurs de relation, présence dans *m* méthodes.
- Gestion des conj. de coordination et ellipses,
- Extraction terminologique utilisant le parallélisme visuel :

## Approches pratiques [ modifier | modifier le code ]

---

L'algorithmique a développé quelques stratégies pour résoudre les problèmes :

- **algorithme glouton** : un premier algorithme peut souvent être proposé en étudiant le problème très progressivement : on résout chaque sous-problème localement en espérant que l'ensemble de leurs résultats composera bien une solution du problème global. On parle alors d'algorithme glouton. L'algorithme glouton n'est souvent qu'une première étape dans la rédaction d'un algorithme plus performant.
- **diviser pour régner** : pour améliorer les performances des algorithmes, une technique usuelle consiste à diviser les données d'un problème en sous-ensembles de tailles plus petites, jusqu'à obtenir des données que l'algorithme pourra traiter au cas par cas. Une seconde étape dans ces algorithmes consiste à « fusionner » les résultats partiels pour obtenir une solution globale. Ces algorithmes sont souvent associés à la récursivité.
- **recherche exhaustive (ou combinatoire)** : une méthode utilisant l'énorme puissance de calcul des ordinateurs consiste à regarder tous les cas possibles. Cela n'est pour autant possible que dans certains cas particuliers (la combinatoire est souvent plus forte que l'énorme puissance des ordinateurs, aussi énorme soit-elle)
- **aléatoire, ou par approximations successives** : certains algorithmes utilisent des recherches aléatoires, ou par approches successives, donnant de meilleurs résultats (en moyenne) que des recherches directes ou explicites.
- **décomposition top-down / bottom-up** : les décompositions top-down consistent à essayer de décomposer le

# Conclusion et Perspectives générales

## Conclusion :

1. Arbre de dépendance du document,
2. Extraction de relations à l'aide de cette structure,
3. Travail en cours avec les structures énumératives.

## Perspectives générales :

- **Extraction de relations :**
  - Traitement du parallélisme visuel,
  - Extension à d'autres objets textuels.
- **Passage à l'échelle sur Wikipedia Fr :**
  - Documents : 1,5 million
  - SE : 3,2 millions
  - Items : 15,6 millions
  - Analyse logique : 3h17 de traitement sur 2,4 Ghz
- **Approche semi-supervisée :**
  - utiliser le corpus LARA comme seed,
  - extraction par domaine (e.g. : vin),
  - extraction « ouverte » (e.g. : sites de laboratoires).



# Exemples supplémentaires

## Paragraphe imbriqué

acteurs régionaux qui semblent les plus à même de peser sur les différentes parties alors que la France peut difficilement jouer un rôle de même nature faute de la panoplie de moyens dont ceux-ci disposent (liens personnels, familiaux, instruments financiers et discrétion liée à la proximité et aux méthodes des régimes non démocratiques).

Elle a néanmoins la possibilité d'accompagner plus activement ce processus régional. C'est ainsi qu'elle peut notamment :

- Appuyer l'action de l'Arabie Saoudite et renforcer la concertation avec Riyad dans le but d'identifier pleinement et en temps voulu les occasions de compromis possibles.
- Engager des contacts avec l'Iran sur le dossier libanais en le dissociant des autres dossiers régionaux. Les hésitations à nouer un dialogue avec celui-ci sur ce dossier méritent d'être réexaminées. Il ne s'agirait pas de le faire en liant les différents dossiers (programme nucléaire, Irak, Palestine et Liban) en vue d'un quelconque « marché » mais sur la base d'une évaluation objective de l'importance des enjeux pour Téhéran dans chacun de ces dossiers. S'il est sûr que l'Iran conservera une position radicale sur son programme nucléaire, il existe de bonnes raisons de penser qu'il est disposé à coopérer pour trouver une solution de compromis au Liban, à la différence de la Syrie.

Le soutien de Téhéran au Hezbollah fait certes partie d'une stratégie iranienne d'appui sur la communauté chiite régionale. Mais la guerre de l'été a conduit le mouvement à payer un prix très élevé et à prendre des risques politiques majeurs qui ébranlent son image de force politique nationale oeuvrant pour son intégration dans le jeu politique libanais sur la base d'un nouveau consensus intercommunautaire. L'Iran ne peut instrumentaliser le Hezbollah que dans certaines limites et risque de susciter des divisions au sein du mouvement comme cela a été le cas dans le passé.

- Obtenir de l'Iran et de l'Arabie Saoudite qu'ils travaillent conjointement pour freiner les appétits et les interventions de Damas au Liban.
- Engager un dialogue avec des représentants du Hezbollah dans un cadre officieux ou à un niveau diplomatique peu élevé pour tester pleinement ses intentions.
- La relation bilatérale avec la Syrie est réduite, pour des raisons compréhensibles, à neutraliser le rôle de Damas au Liban et il est difficile d'envisager un réel réchauffement avec les Syriens avant la mise en place du tribunal à caractère international que prévoit la résolution 1701 du Conseil de Sécurité qui nous engage d'autant plus que nous en avons

# Exemples supplémentaires

## Citation avec mise en forme de paragraphe

### 1 Introduction

Les travaux descriptifs qui sont faits sur les propositions subordonnées circonstancielles concernent essentiellement celles qui ont un verbe conjugué à un mode personnel. Ils font abstraction des subordonnées participiales, (désormais SP) qui sont reconnues dans l'exemple suivant : *le chat parti, les souris dansent*. Beaucoup de grammaires de référence ignorent cette construction (Wagner et Pinchon 1991), d'autres la méconnaissent (Wilmet, 1997). Celles qui l'évoquent l'expliquent en la mettant en équivalence avec une subordonnée circonstancielle conjonctive en *dès que* ou *lorsque* : *dès que le chat est parti, les souris dansent* (Grevisse 1993 ; Riegel et al 1994).

Dans leur analyse, les grammaires qui parlent de la construction avancent l'idée que son procès est antérieur à celui de la proposition qui l'héberge (désormais PH), en précisant que la construction peut être indifféremment précédée par des éléments comme *une fois, sitôt, aussitôt*, et que le participe peut être précédé de l'auxiliaire *étant*. C'est également la position d'A. Borillo (2006 : 5) dans son étude sur les structures participiales à prédication seconde. Elle avance l'explication suivante :

« On peut constater que l'absence du marqueur temporel est parfois possible, sans réelle modification du sens de l'énoncé, si ce n'est que *sitôt, aussitôt*, et *à peine* ajoutent effectivement une précision d'immédiateté et que *une fois* souligne de manière explicite la relation d'antériorité d'une première éventualité par rapport à une autre. *Une fois le texte rédigé, il fallut le taper sur un stencil* ; *le texte rédigé, il fallut le taper sur un stencil*. Sans marqueur temporel, le sens reste très proche, de même que les règles de construction : le participe passé est celui d'un verbe construit avec le verbe *être*, qui doit être interprété avec une valeur passive si le verbe est transitif, avec une valeur active si le verbe est inaccusative ».

L'objectif de mon propos est de montrer que, pour mieux comprendre le fonctionnement de la SP, celle-ci doit être analysée, non pas dans le cadre de la phrase, mais dans le cadre du discours. En effet, comme B. Combettes (1993) l'a montré, les constructions détachées sont des éléments qui assurent la continuité thématique du discours. Elles reprennent, en général, des référents contenus dans le contexte antérieur. En tant que telle, la SP peut difficilement avoir un référent nouveau. Elle a en général un référent qui est déjà présent dans le discours. Il paraît donc difficile de se contenter de l'analyse phrastique pour rendre compte de ce type de construction. B. Combettes (1993 : 39-40) l'a souligné : « La construction détachée apparaît [...] comme un constituant dont le fonctionnement dépend autant, sinon plus, de contraintes textuelles, de facteurs discursifs, que de caractéristiques strictement syntaxiques : le prédicat réduit qu'elle constitue se comporte en fait comme un prédicat intermédiaire, passage entre deux énoncés, qui prolonge le contexte de gauche dans une fonction de maintien d'un référent thématique. »

# Exemples supplémentaires

## Items avec mise en forme de paragraphe

### 3.4 L'analogie en marge du générativisme

Même aux beaux jours de la grammaire générative, l'analogie en tant que processus morphologique conservait cependant des défenseurs, d'autant plus virulents parfois que le phénomène se trouvait marginalisé :

– Motsch (1987 : 24) se demande ainsi s'il est fondé d'opposer analogie et règles. Comme d'autres avant (par ex. van Marle, 1985) et après lui (par ex. Biela, 1991 : 114-5), il souligne en effet que les règles n'existent qu'en tant qu'elles sont incarnées par des mots existants, présentant des similarités :

The creation of new words (...) presupposes rules. But rules need not have an existence of their own. We may conceive of rules as the result of a process of analysis operating on similarity of item of the vocabulary.

– T. Becker va plus loin. Non seulement il considère que les règles constituent des abstractions faites à partir de paires de mots existants (en 1990, il écrit que toutes les règles sont des analogies), mais encore il fait l'hypothèse en 1993 de deux types de morphologie, orthogonaux l'un à l'autre. Du point de vue de la morphologie qu'il appelle 'syntagmatique', qui est celui du linguiste, les mots construits peuvent être décrits comme des agencements de morphèmes (1993 : 1) ; du point de vue de la morphologie dite 'paradigmatique', qui est celui du locuteur, ils n'ont pas de structure, ce sont des « signes minimaux » (2003 : 272).

– La perspective de R. Skousen, seul ou en collaboration, est différente. Après avoir montré en 1989 l'incapacité des systèmes basés sur règles à venir à bout des règles non-déterministes et proposé une définition mathématique de l'analogie, il décrit en 1992 un algorithme exploitant les similarités entre mots attestés pour prédire quelle forme revêtira un mot nouveau.

Pour des raisons que développe van Marle (2000 : 226-*sq.*), la question de l'analogie est souvent polémique, et les positions prises à son égard sont la plupart du temps extrêmes : on peut de la sorte reprocher leur manichéisme à Derwing & Skousen (1989), nettement en faveur de l'analogie, comme on peut reprocher le sien à Plag (1999), partisan, lui, d'une morphologie basée sur des règles. En témoignent les vives critiques qu'ont suscitées les travaux de Becker et le modèle analogique de Skousen (*cf.* entre autres Bauer, 1993, Baayen, 1995, et Plag, 1999)<sup>15</sup>. En substance, il est reproché à l'analogie :

- de ne pas permettre de bonnes prédictions sur les formes possibles et impossibles,
- d'être insuffisamment contrainte,
- de ne pas permettre de généralisation,
- d'avoir contre elle des évidences psychologiques,

# Exemples supplémentaires

## Imbrication d'objets textuels

### 5.1.3 Le lexique.

La grammaire traditionnelle adopte une attitude ambiguë au regard du lexique. D'une part, on considère qu'il est connexe au domaine qu'elle couvre, et ne concerne donc pas au premier chef le grammairien, mais d'autre part, l'étude des registres de langage (familier, littéraire...), qui passe par le lexique, est généralement envisagée dans les grammaires actuelles. Si l'on veut cependant favoriser l'acquisition de routines de compréhension et de production de textes – en incitant les apprenants à mettre en corrélation sémantique des faisceaux d'indices congruents – il est alors nécessaire de relier l'emploi du lexique à celui des “outils grammaticaux” traditionnels, en relation aussi au domaine textuel (cf. Barbazan 2007b pour une application dans un objectif de didactique en FLE à l'emploi de temps verbaux).

Par ailleurs (ainsi que nous l'avons suggéré au point 3.1.), les conclusions de l'étude de Kerbrat-Orecchioni (1997) vont dans le sens de l'inscription effective d'une dimension énonciative au sein du signifié d'une catégorie lexicale, marquée par un trait [+subjectif], la catégorie des *subjectivèmes*. Cette dimension énonciative est conjointe à la dimension dénotative (référentielle).

« Ces substantifs cumulent deux types d'informations, d'ailleurs indissociables :

- une description du dénoté
- un jugement évaluatif, d'appréciation ou de dépréciation, porté sur ce dénoté par le sujet d'énonciation. » (Kerbrat-Orecchioni 1997, 73)

Entre autres qualités, que nous ne pouvons pas reprendre ici, ces termes

« sont à éliminer d'un discours à prétention d'objectivité, dans lequel le locuteur refuse de prendre position par rapport au dénoté évoqué. [C'est pourquoi ils] peuvent être considérés comme comportant un trait sémantique [+subjectif] » (Kerbrat-Orecchioni 1997, 73).

Logiquement, et en corrélation avec l'adoption de ce trait [+subjectif] pour certains termes (toutes catégories lexicales confondues), on peut prévoir un trait [-subjectif] pour d'autres. Ces derniers, souvent décrits comme “neutres”, alors qu'ils sont aussi dénotatifs d'une attitude énonciative que les *subjectivèmes*, sont privilégiés par exemple dans les rapports de police. On voit par cet exemple se profiler la possibilité de mettre en relation la “couleur énonciative” d'un terme lexical avec un mode de textualisation privilégié, en relation avec la caractérisation des genres. Il faut bien sûr se méfier ici de la caricature descriptive, péchant par excès de systématisation et source de surgénéralisations inévitables pour les apprenants.

# Exemples supplémentaires

## Parallélisme visuel et lexical

### 4.2 Les données : paramètres morphophonologiques

Si l'on y regarde de plus près, on constate que la forme du dérivé dépend principalement de celle de l'adjectif (qui est aussi, rappelons-le, celles du nom de personne et du nom de langue). Trois ensembles se dessinent (on laissera de côté les finales rarement représentées et les exceptions, qui ne remettent pas en cause le classement proposé) :

– Si l'adjectif a une finale en *-al* (*provençal*), *-an* (*andorran*), *-ain* (*américain*), *-in* (*latin*), *-on* (*gascon*) ou une finale non suffixoïde (*arabe*, *berbère*, *corse*...), la concaténation se fait, sauf exception, sur cette forme : *provençalisme*, *andorranisme*, *américanisme*, *latinisme*, *gasconnisme*, *arabisme*, *berbérisme*, *corsisme*...

– Si l'adjectif a une finale suffixale ou suffixoïde en *-ien* (*italien*, *autrichien*, *palestinien*), *-éen* (*européen*), *-ique* (*attique*, *gaélique*), *-and* (*allemand*, *romand*), *-ard* (*picard*), *-ol* (*espagnol*), la concaténation se fait tantôt sur cette forme, moyennant le cas échéant une allomorphie (*italianisme*, *européanisme*, *atticisme*, *romandisme*, *picardisme*, *espagnolisme*...), tantôt sur une forme amputée de sa dernière rime (*palestinisme*, *européisme*, *gaélisme*...), tantôt sur une base supplétive (éventuellement tronquée) (*austriacisme*, *germanisme*, *hispanisme*...). Les doublets sont relativement nombreux, plusieurs solutions pouvant être exploitées pour une même base.

– Si l'adjectif a une finale suffixale ou suffixoïde en *-ais* (*anglais*, *japonais*, *libanais*, *portugais*...) ou en *-ois* (*hongrois*, *chinois*, *québécois*...), la concaténation est exceptionnelle. Les dérivations se font sur une base tronquée ou sur le nom de pays (*japonisme*, *libanisme*, *québécoisme*...), ou bien sur une base supplétive (*anglicisme*, *lusitanisme*, *magyarisme*...). Les hésitations entre plusieurs formes, ici encore, sont nombreuses. Pour *Chine* / *chinois*, on en trouve quatre – (*anti*-)*chinoisisme*, (*anti*-)*chinisme*, *sinisme*, *sinicisme* – dont aucune n'est réellement usuelle, ce qui traduit sans doute une situation de blocage.

- S. Afantenos, P. Denis, P. Muller, and L. Danlos. Learning recursive segments for discourse parsing. In *Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010)*, 2010.
- S. D Afantenos, N. Asher, F. Benamara, M. Bras, C. Fabre, M. Ho-Dac, A. Le Draoulec, P. Muller, M.-P. Péry-Woodley, L. Prévot, Rebeyrolles J., Tanguy T., Vergez-Couret M., and Vieu L. An empirical resource for discovering cognitive principles of discourse organisation : the annodis corpus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, pages 2727–2734, 2012.
- Sophie Aubin and Thierry Hamon. Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, pages 380–387. Springer, 2006.
- N. Aussenac-Gilles and M.-P. Jacques. Designing and evaluating patterns for relation acquisition from texts with caméléon. *Terminology*, 14 :45–73, 2008.
- Matthew Berland and Eugene Charniak. Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 57–64. Association for Computational Linguistics, 1999.
- C. Bush. Des déclencheurs des énumérations d'entités nommées sur le web. *Revue québécoise de linguistique*, 32(2) :47–81, 2003.

# Références II

- F. Y. Y Choi. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 26–33. Association for Computational Linguistics, 2000.
- Anne Condamines and Josette Rebeyrolle. Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. In *Actes Journées Ingénierie des Connaissances et Apprentissage Automatique (JICAA'97)*, pages 191–206, 1997.
- Béatrice Daille. Study and implementation of combined techniques for automatic extraction of terminology. *The balancing act : Combining symbolic and statistical approaches to language*, 1 :49–66, 1996.
- J.L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5) :378, 1971.
- M. A. K. Halliday and R. Hasan. *Cohesion in English*. Longman, London, 1976.
- M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics*, volume 2, pages 539–545. Association for Computational Linguistics, 1992.
- N. Hernandez and B. Grau. Détection automatique de structures fines de texte. In *Actes de la 12e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2005)*, 2005.

# Références III

- L.-M. Ho-Dac, M.-P. Péry-Woodley, and L. Tanguy. Anatomie des structures énumératives. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2010)*, 2010.
- Nanda Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 22. Association for Computational Linguistics, 2004.
- M. Kamel and B. Rothenburger. Elicitation de structures hiérarchiques à partir de structures énumératives pour la construction d'ontologie. In *Journées Francophones d'Ingénierie des Connaissances (IC 2011)*, pages 505–522, Annecy, 2011.
- C. Luc. Une typologie des énumérations basée sur les structures rhétoriques et architecturales du texte. In *Actes de la 8e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2001)*, pages 263–272, 2001.
- Christophe Luc. *Représentation et composition des structures visuelles et rhétoriques du textes. Approche pour la génération de textes formatés*. PhD thesis, Université Paul Sabatier, 2000.
- W.C. Mann and S.A. Thompson. Rhetorical structure theory : Toward a functional theory of text organization. *Text*, 8(3) :243–281, 1988.
- D. Marcu. Automatic discourse parsing. In K. Brown, editor, *Encyclopedia of Language and Linguistics*. Elsevier, 2nd edition, 2006.

# Références IV

- Y. Mathet and A. Widlöcher. Une approche holiste et unifiée de l'alignement et de la mesure d'accord inter-annotateurs. In *Actes de la 18e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2011)*, 2011.
- Eugene W Myers. An  $O(n^2)$  difference algorithm and its variations. *Algorithmica*, 1 (1-4) :251–266, 1986.
- Adeline Nazarenko. Sur quelle sémantique reposent les méthodes automatiques d'accès au contenu textuel? *Sémantique et corpus*, pages 211–244, 2005.
- J. Pustejovsky and A. Stubbs. *Natural language annotation for machine learning*. O'Reilly, 2012.
- Marie-Paule Péry-Woodley and Donia Scott. Computational approaches to discourse and document processing. *TAL*, 47(2) :7–19, 2006.
- C. Ramakrishnan, A. Patnia, E. H. Hovy, and G. Burns. Layout-aware text extraction from full-text pdf of scientific articles. *Source code for biology and medicine*, 7(1), 2012.
- Barbara Rosario and Marti A Hearst. Classifying semantic relations in bioscience texts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 430. Association for Computational Linguistics, 2004.
- J. Virbel. Structures textuelles, planches fascicule 1 : Enumérations, version 1., Technical report, IRIT, 1999.
- Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3 :1083–1106, 2003.