La mise en forme des textes : un indice supplémentaire pour l'identification des relations sémantiques

Jean-Philippe Fauconnier

Institut de Recherche en Informatique de Toulouse

Marseille
20 mai 2016

Positionnement théorique

Relations sémantiques

- Utiles pour la représentation des connaissances au travers de ressources :
  - p. ex. WordNet (Fellbaum, 1998)
  - p. ex. DBpedia (Auer et al., 2007)
- Rôle crucial dans des applications de plus haut niveau

Distinction relations paradigmatiques vs. syntagmatiques

- **Paradigmatique** : regroupent les lexies d'un même paradigme constitutives de classes sémantiques ou directement reliées sémantiquement.
  p. ex. voiture et véhicule
- **Syntagmatique** : dénotent des capacités d'association d'une lexie donnée et du contexte que celle-ci sélectionne.
  p. ex. chaperon et rouge

Positionnement théorique

Objet de ce travail

Extraction de relations paradigmatiques entre entités textuelles (termes et entités nommées) à partir de textes en langage naturel

Approche pionnière (Hearst, 1992)

- Utilisation de patrons lexico-syntaxiques
- Expression linguistique de l'hyperonymie
- Reprise et adaptée par la suite (Morin, 1999; Séguéla, 2001)

SN (tel(le)?(s)? que|comme) (SN (,)?)* (ou|et)? SN

En outre, des animaux tels que l'éléphant, le cheval, le lama ou le chat peuvent être domptés et dressés.

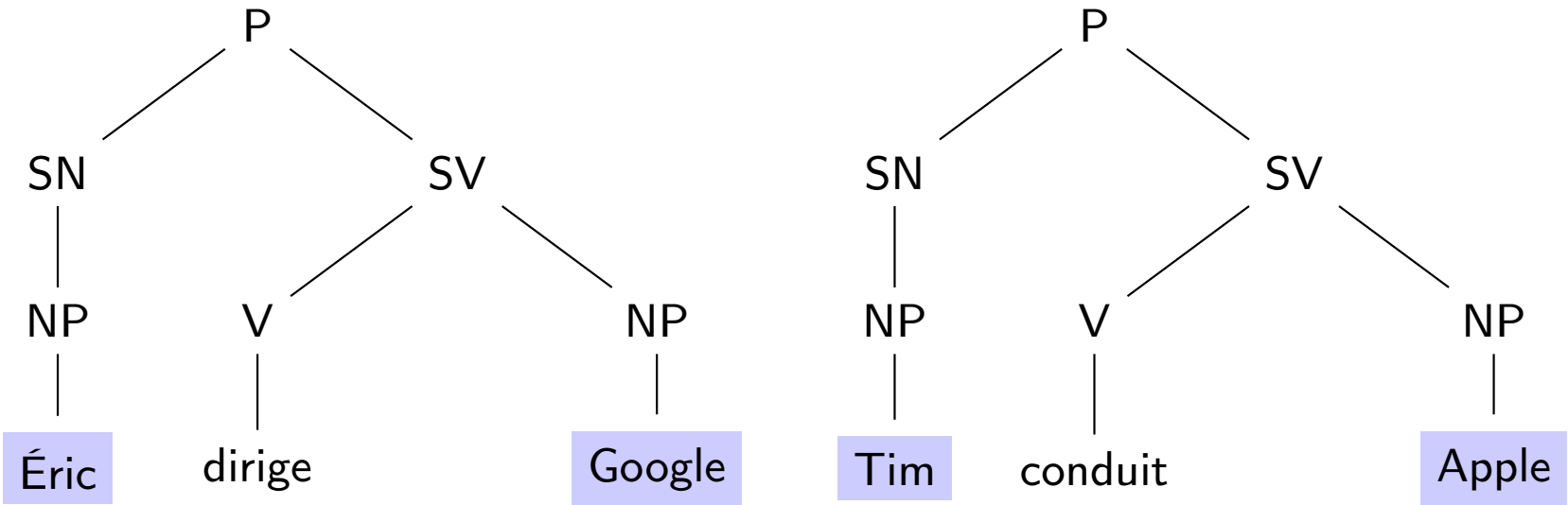
Positionnement théorique

Approches suivantes

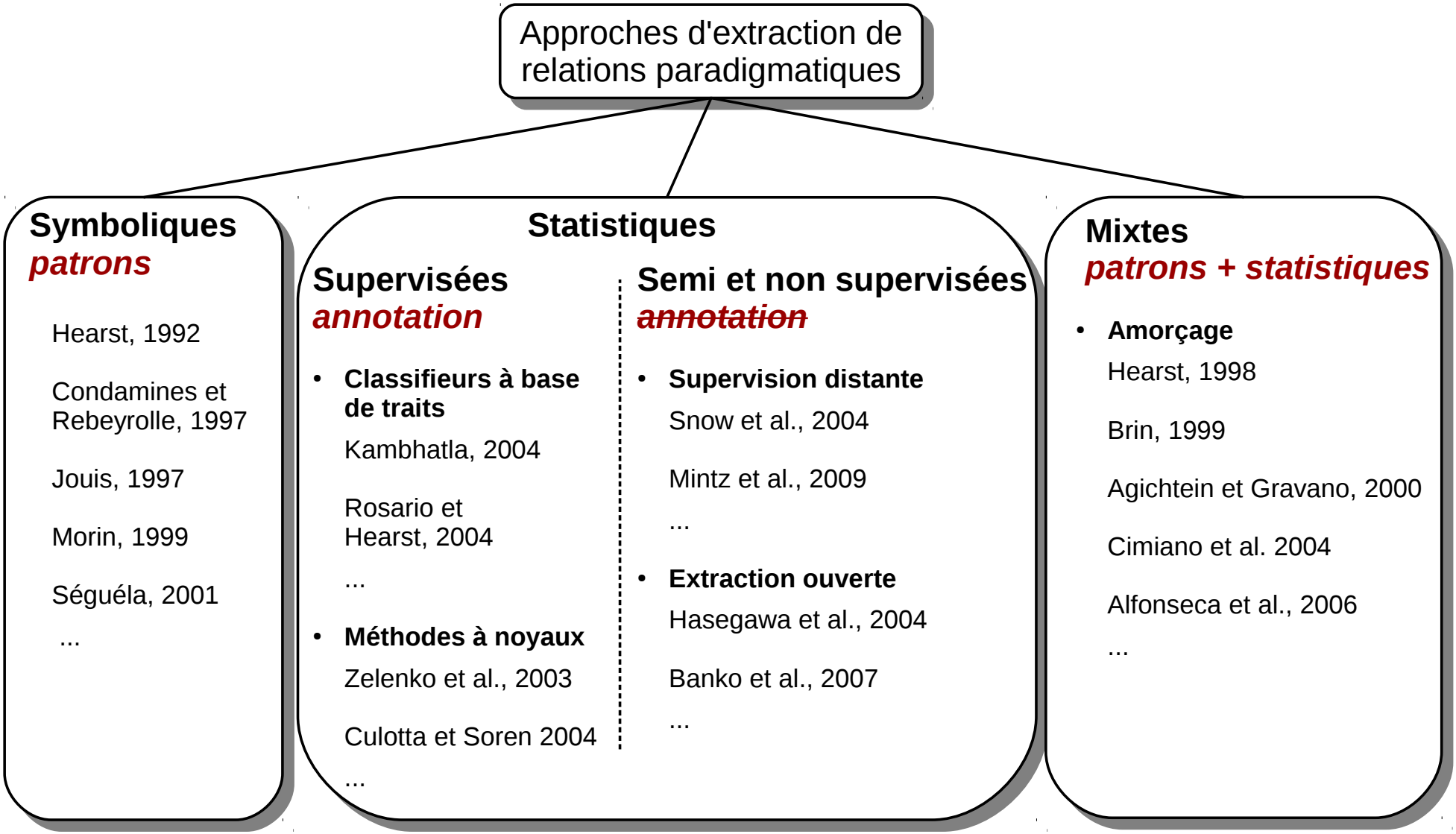
- Popularisation des méthodes statistiques
- Parallèles à l'essor des campagnes MUC et ACE
- Autres types de relations (p. ex. localisation, affiliation, etc.)

P. ex. méthodes à noyaux pour la similarité entre arbres (Zelenko et al., 2003)

$$f(T) = \text{sgn}\Big(\sum_{i=1}^{l} \alpha_i y_i \; K(T_i, T) + b\Big)$$



Positionnement théorique



Mise en forme et extraction de relations

Problématique

- Les approches présentées se limitent généralement au niveau phrastique.
- Or, il est possible d'exploiter la mise en forme pour découvrir de nouvelles relations au-delà des frontières de la phrase.

Les indices typographiques et dispositionnels signalent, au même titre que des indices lexico-syntaxiques avec lesquels ils se combinent, des phénomènes sémantiques à l'échelle du texte, incluant l'expression de relations sémantiques.

Structures énumératives verticales

Terrain idéal pour l'étude des interactions entre mise en forme et richesse sémantique (Porhiel, 2007) :

- Elles sont porteuses de marques qui facilitent leur repérage
- Elle sont propices à la présence de relations sémantiques

Exemple de structure énumérative verticale

Dès qu'un port atteint une taille suffisante, un certain nombre de navires de services y sont basés ; ils ne font pas partie du trafic du port mais sont utilisés pour différentes opérations portuaires. On trouve ainsi :

- Les dragues, de différents types suivant la nature du fond et la zone à couvrir (à élinde traînante, à godets. . .) ; elles servent à maintenir une profondeur suffisante dans le port et les chenaux d'accès, malgré l'apport de sédiments dû aux rivières et courants.
- Les bateaux pilote servant à amener les pilotes à bord des navires de commerce arrivant au port. Sur les ports de moyenne importance, on trouve quelques pilotines opérant à partir du port ; sur les grands ports de commerce, on trouve parfois un grand navire dans la zone d'atterrissage hébergeant les pilotes, et duquel partent les pilotines.
- Les remorqueurs portuaires qui servent à aider les grands navires à manœuvrer durant les opérations d'amarrage et d'évitage.
- Les bateaux de lamanage utilisés par les lamaneurs pour porter les amarres à terre.

Exemple de structure énumérative verticale

amorce

Dès qu'un port atteint une taille suffisante, un certain nombre de navires de services y sont basés ; ils ne font pas partie du trafic du port mais sont utilisés pour différentes opérations portuaires. On trouve ainsi :

énumération

item
- Les dragues, de différents types suivant la nature du fond et la zone à couvrir (à élinde traînante, à godets. . .) ; elles servent à maintenir une profondeur suffisante dans le port et les chenaux d'accès, malgré l'apport de sédiments dû aux rivières et courants.

item
- Les bateaux pilote servant à amener les pilotes à bord des navires de commerce arrivant au port. Sur les ports de moyenne importance, on trouve quelques pilotines opérant à partir du port ; sur les grands ports de commerce, on trouve parfois un grand navire dans la zone d'atterrissage hébergeant les pilotes, et duquel partent les pilotines.

item
- Les remorqueurs portuaires qui servent à aider les grands navires à manœuvrer durant les opérations d'amarrage et d'évitage.

item
- Les bateaux de lamanage utilisés par les lamaneurs pour porter les amarres à terre.

Exemple de structure énumérative verticale

Dès qu'un port atteint une taille suffisante, un certain nombre de navires de services y sont basés ; ils ne font pas partie du trafic du port mais sont utilisés pour différentes opérations portuaires. On trouve ainsi :

- Les dragues, de différents types suivant la nature du fond et la zone à couvrir (à élinde traînante, à godets. . .) ; elles servent à maintenir une profondeur suffisante dans le port et les chenaux d'accès, malgré l'apport de sédiments dû aux rivières et courants.
- Les bateaux pilote servant à amener les pilotes à bord des navires de commerce arrivant au port. Sur les ports de moyenne importance, on trouve quelques pilotines opérant à partir du port ; sur les grands ports de commerce, on trouve parfois un grand navire dans la zone d'atterrissage hébergeant les pilotes, et duquel partent les pilotines.
- Les remorqueurs portuaires qui servent à aider les grands navires à manœuvrer durant les opérations d'amarrage et d'évitage.
- Les bateaux de lamanage utilisés par les lamaneurs pour porter les amarres à terre.

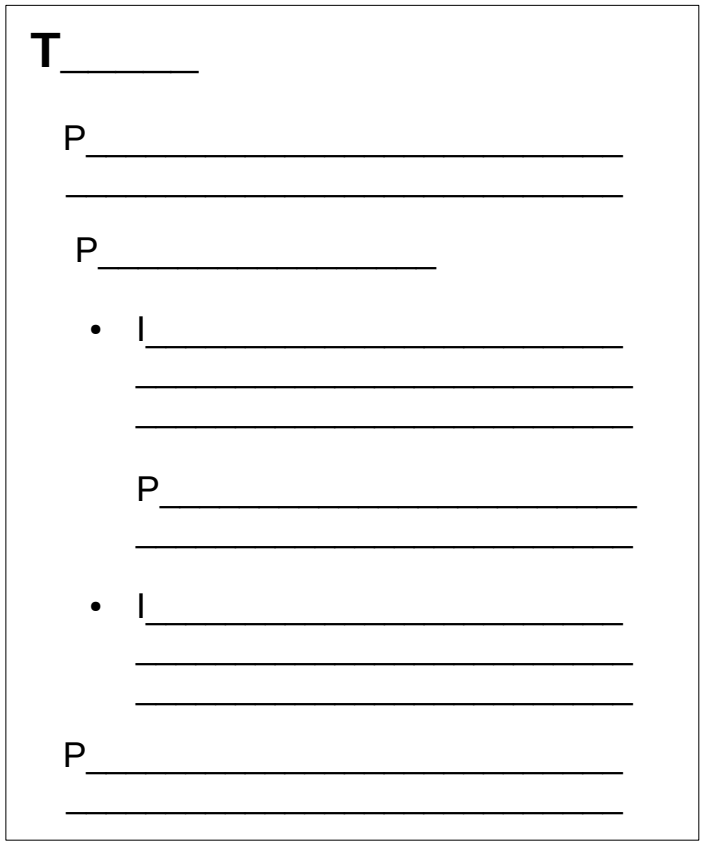
Plan

Partie I

Structure de Document

Notion de document textuel

Image de texte (Pascual, 1991)



3 niveaux de structuration

(1) **Structure visuelle** forme dans laquelle apparaît un document
ex. bloc textuel, figure, etc.

(2) **Structure logique** niveau ordonnant les unités logiques
ex. paragraphe, titre, etc.

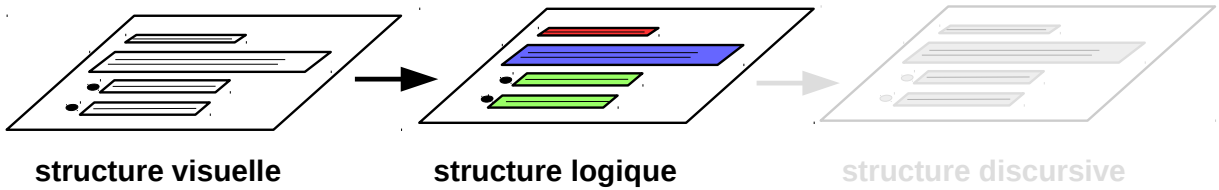
(3) **Structure discursive** organisation du message
ex. introduction, résumé, etc.

- Difficulté à distinguer des frontières nettes
- Interactions dans la construction du sens

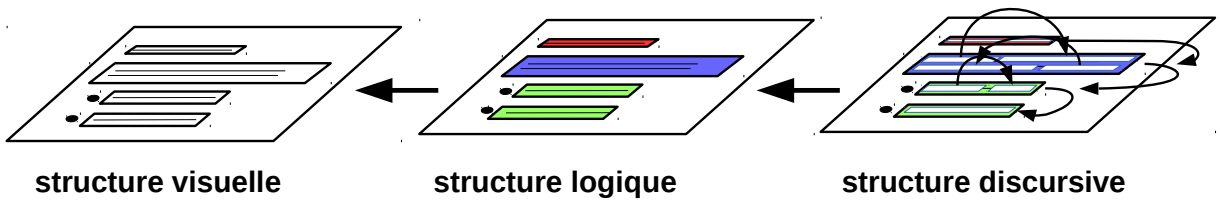
Contexte du travail

Deux communautés scientifiques

- **Analyse du Document (AD) :**
  - Dichotomie uniquement entre structures visuelle et logique
    (Furuta et al., 1982; André et al., 1989, 1990)
  - Intérêt pour l'analyse géométrique de documents complexes

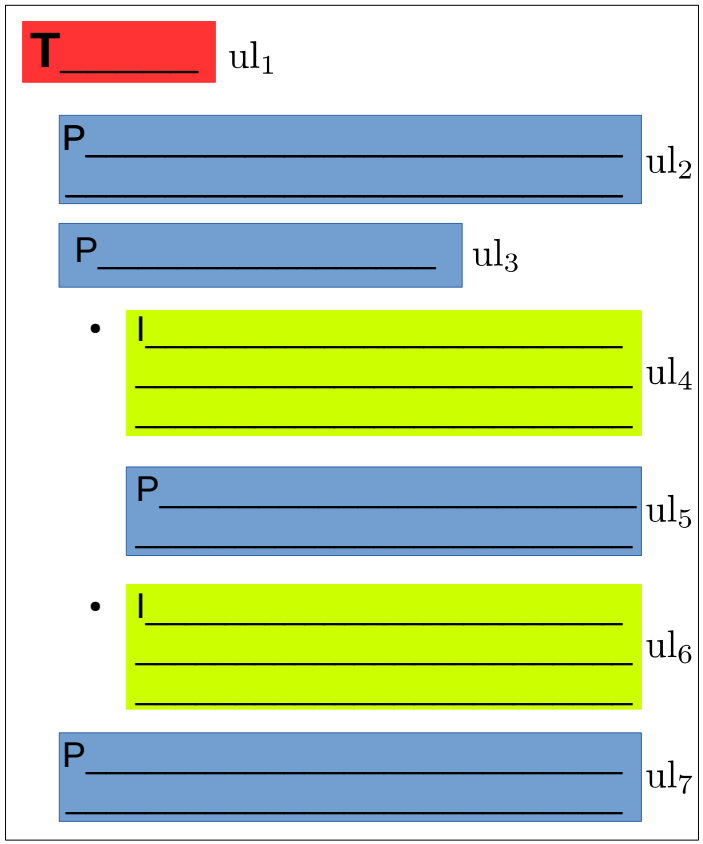


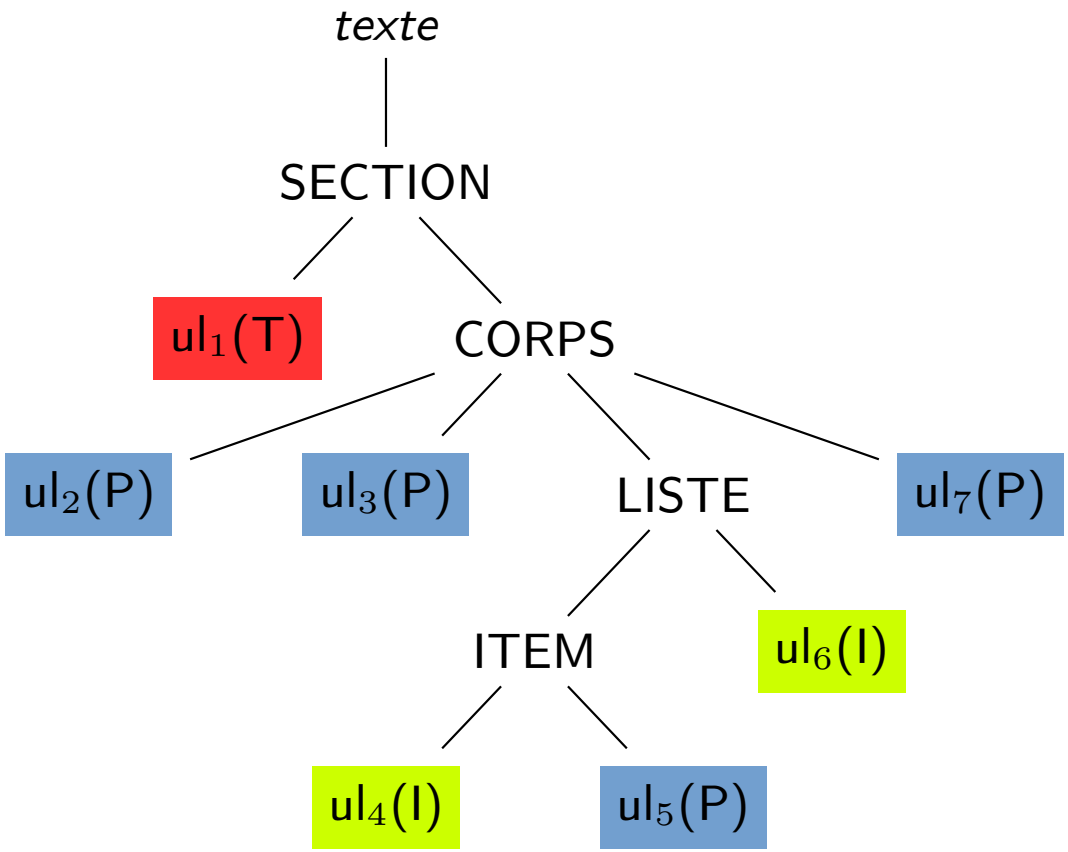
- **Traitement Automatique du Langage (TAL) :**
  - Distinction entre structures visuelle, logique et discursive
    (Virbel, 1989; Bateman et al., 2001; Power et al., 2003)
  - Intérêt pour la génération de texte et sa mise en forme



Modèles TAL

Représentation de la structure logique dans les modèles TAL





- Les unités logiques suivent un principe de composition
- Définition nécessaire de catégories abstraites

⇒ Contraintes difficiles à intégrer dans un processus d'analyse

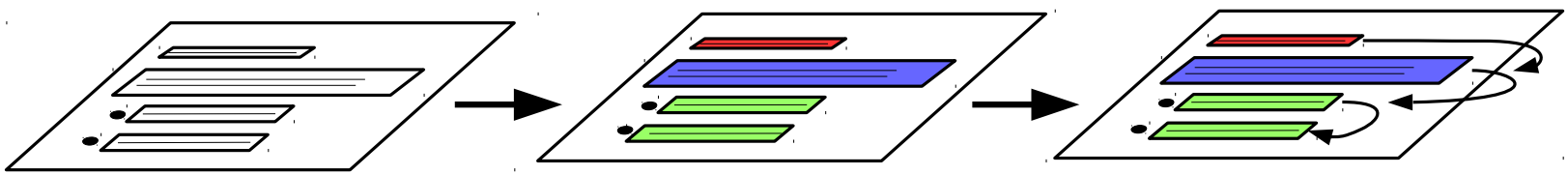
Proposition d'un modèle

Objectif de notre approche

Représenter la structure logique des documents textuels afin d'identifier des structures textuelles hiérarchiques indépendamment de leur marquage typographique et dispositionnel.

Positionnement orthogonal à l'AD et au TAL

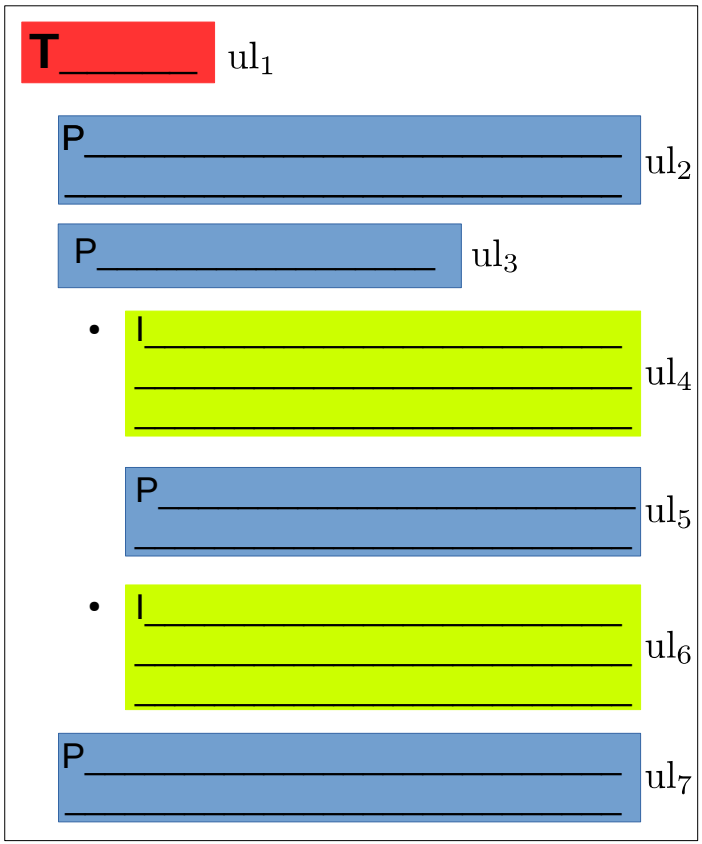
- Processus d'analyse et non de génération de texte
- Abstraction de la mise en forme
- Connexion avec la structure discursive

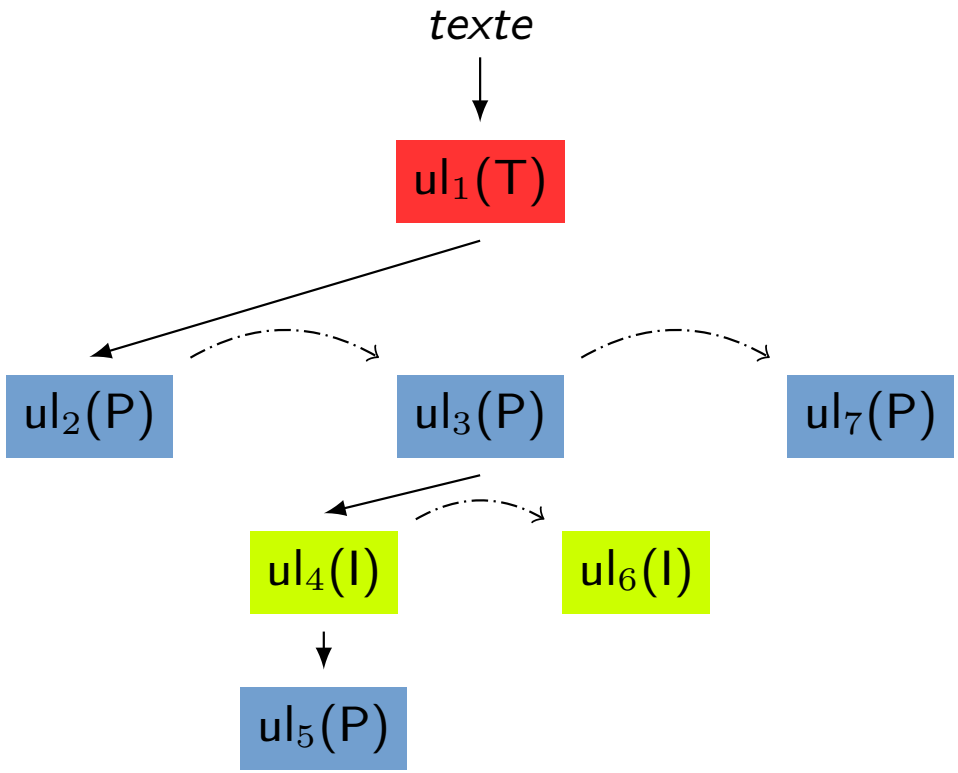


structure visuelle → **structure logique de surface** → **structure logique profonde**

Proposition d'un modèle

Structure logique selon un principe de dépendance





- Les unités logiques suivent un principe de dépendance
- Contraintes syntaxiques : projectivité et transitions à droite

⇒ Processus d'analyse comparable à l'analyse syntaxique en dépendances

Proposition d'un modèle

Deux types de relations de dépendances

Parallèle avec théories du discours (Mann et Thompson, 1988; Asher, 1993)

- **Subordination** Lie des unités logiques de premier plan avec des unités logiques de second plan.
- **Coordination** Lie des unités logiques de même plan (au sens étendu : coordination, juxtaposition).

$\Rightarrow$ Le choix s'effectue sur la présence d'indices de cohésion

Jeu d'étiquettes logiques

- Décomposition de la structure en un ensemble restreint d'étiquettes
- Objets complexes construits par articulation d'unités logiques atomiques

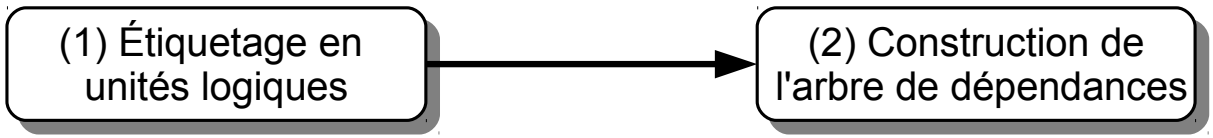
Plan

Identification de la structure de document

Objectif

Identifier automatiquement la structure logique des documents selon le modèle de structuration de document proposé.

Deux tâches



Expériences menées pour le format PDF (TALN, 2014)

- Enrichissement de sous-ensembles du corpus ANNODIS :
  698 pages au total
  LING : articles scientifiques formatés et structurés
  GEOP : rapports et articles sans consensus visuel
- Prétraitement avec LAPDF-Text (Ramakrishnan et al., 2012)

Prétraitement : segmentation en blocs textuels

3.1.2 L'inversion du sujet nominal

Il s'agit de diverses constructions à sujet nominal (non clitique) accordé au verbe qui le précède. Cette inversion a été nommée "inversion stylistique" (Kayne 1973). Il existe de nombreuses études depuis une trentaine d'années sur ce sujet. L'article de Kampers-Mahne, Marandin, Drijkoningen, Doetjes & Hulk (2004) distingue plusieurs types:

1. L'inversion dans les contextes d'extraction (questions partielles, relatives, clivées):

(43) Où est allée Marie? Je me demande où est allée Marie.

(44) La personne qu'a rencontrée Pierre est ma cousine.

(45) C'est dans cette maison qu'est né Vict

2. L'inversion inaccusative (Marandin 2003), avec auxiliaire *être*, passifs); elle est observab

-complétives:

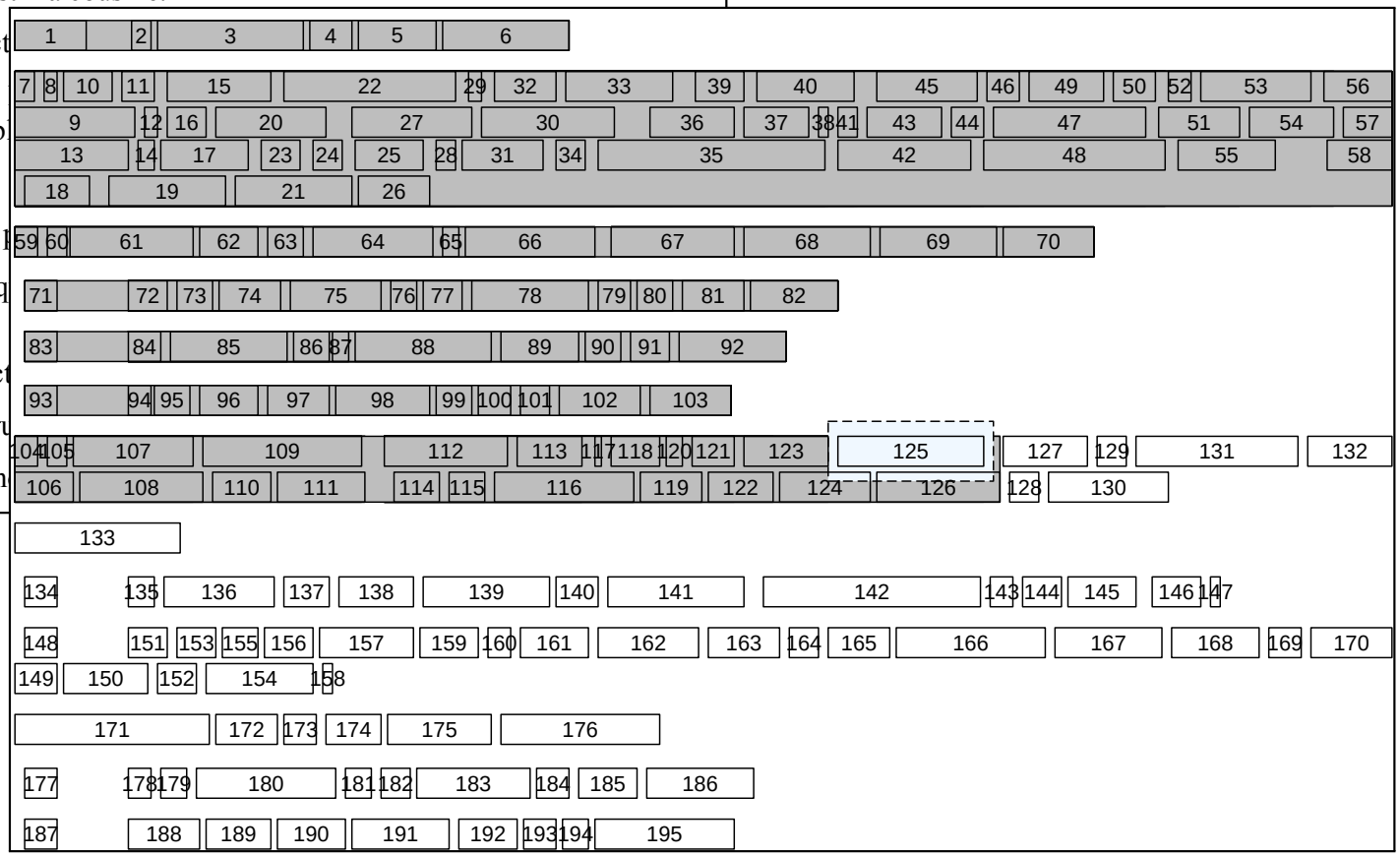
(46) Je voudrais que soient distribués ces

(47) On eût dit que traînait dans la pièce q des syrtes, 32, Frantext).

-indépendantes avec ou sans adverbe introduct

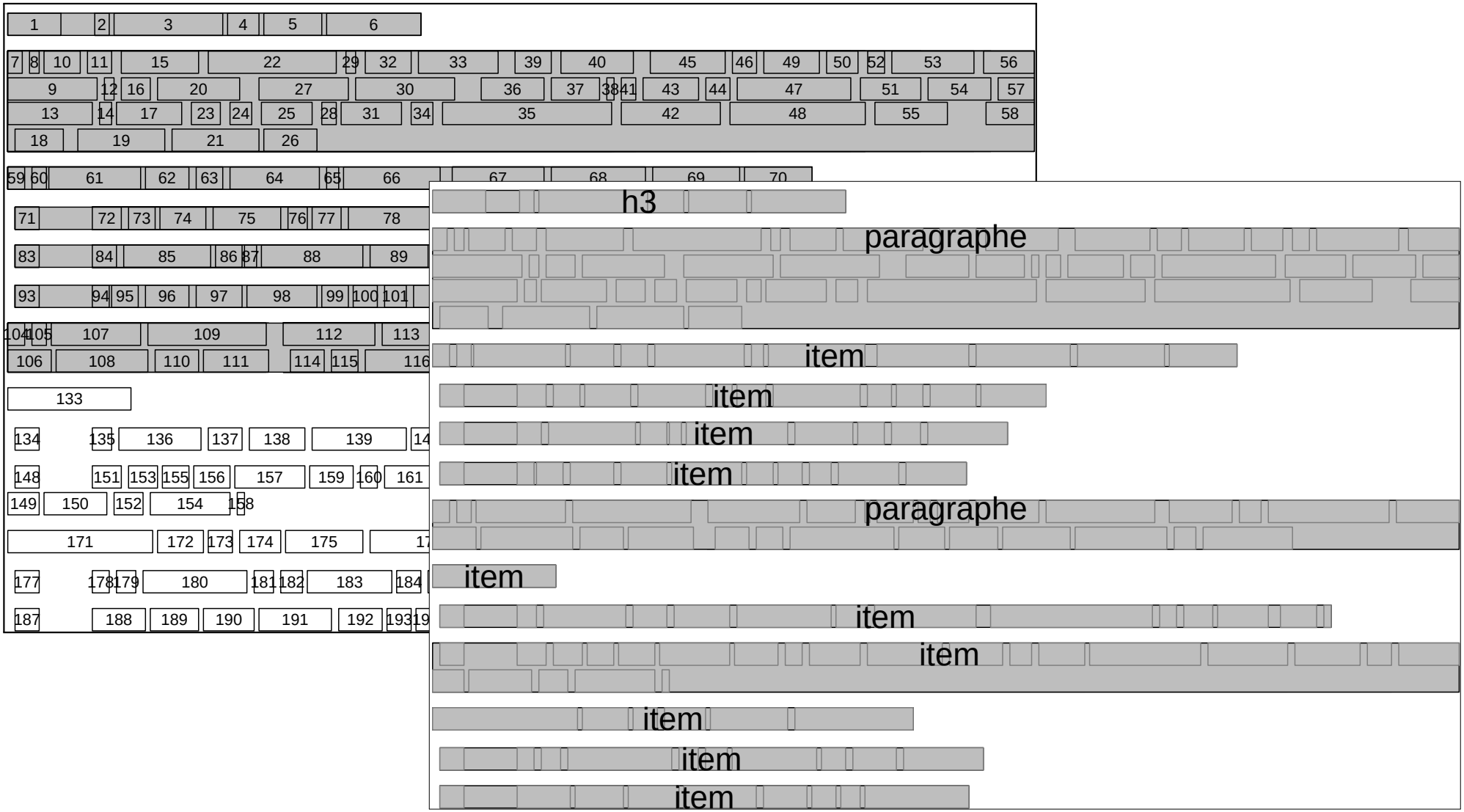
(48) A ce moment-là se fit entendre un bru

(49) Entre alors notre gardien avec de la n



(1) Étiquetage en unités logiques

Enrichissement des blocs avec des étiquettes logiques



(1) Étiquetage en unités logiques

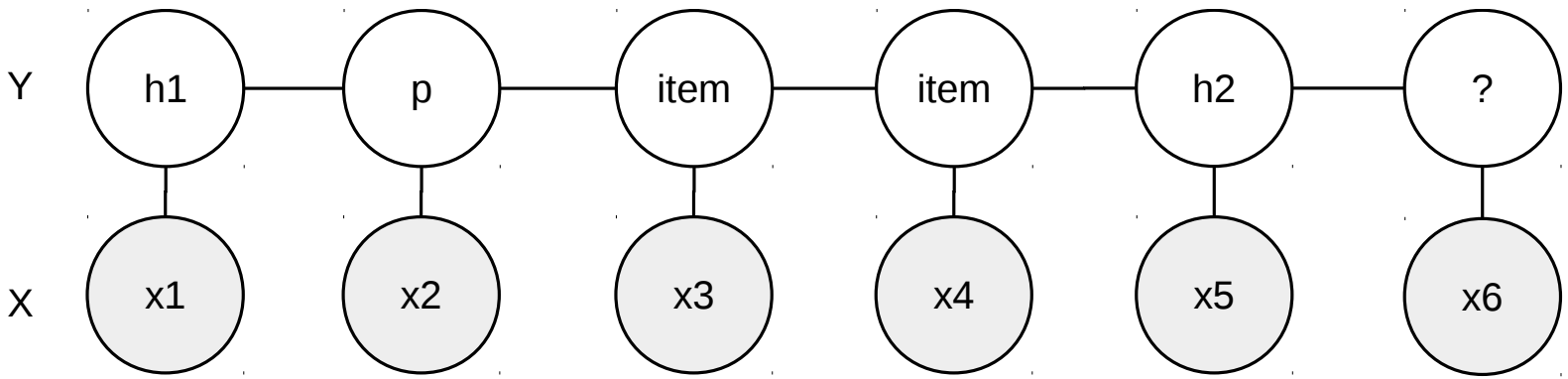
Apprentissage séquentiel

Généralisation statistique de séquences d'unités logiques étiquetées

- Chaque unité logique est représentée par :
  - **Traits d'états** : indentation visuelle, marque d'emphase, etc.
  - **Traits de transitions** : bigrammes d'étiquettes, rupture de fontes, etc.

- Classifieur : Champs Conditionnels Aléatoires

$$p(y_1, y_2, \ldots, y_m | x_1, \ldots, x_m) = \frac{\exp(\theta^\mathsf{T} F(\mathbf{x}, \mathbf{y}))}{\sum_{\mathbf{y}' \in \mathcal{Y}^m} \exp(\theta^\mathsf{T} F(\mathbf{x}, \mathbf{y}'))}$$

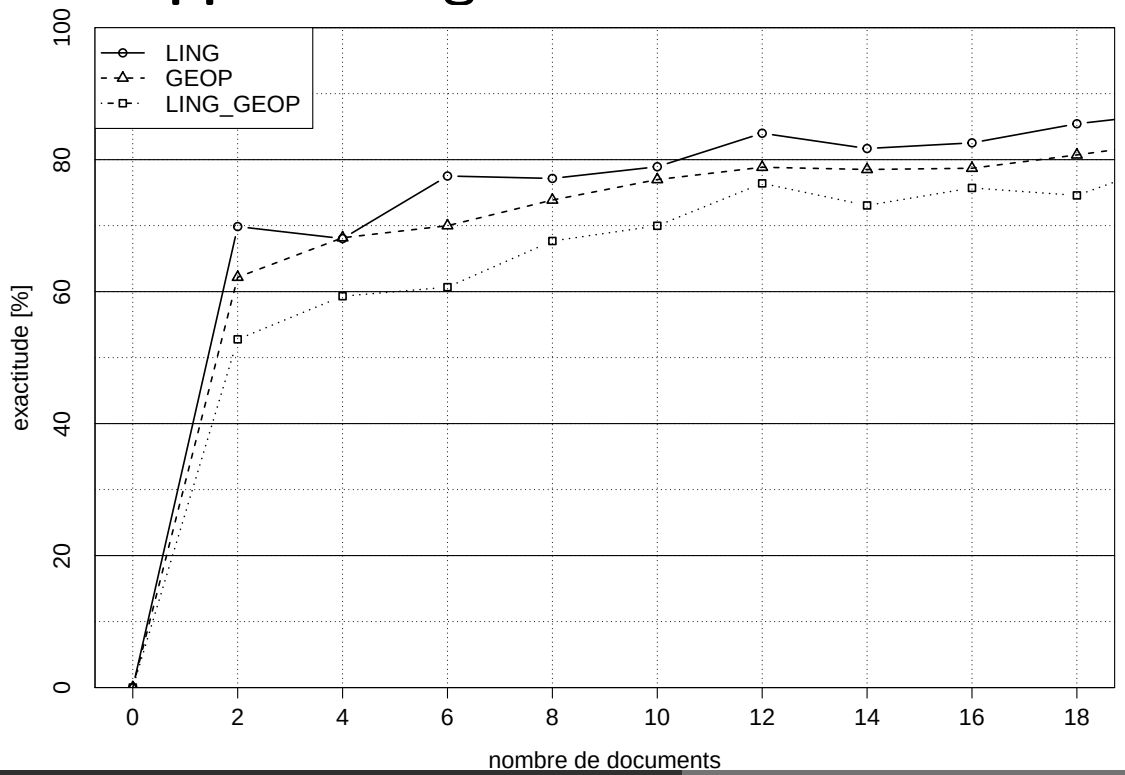


(1) Étiquetage en unités logiques

Évaluation pour l'étiquetage

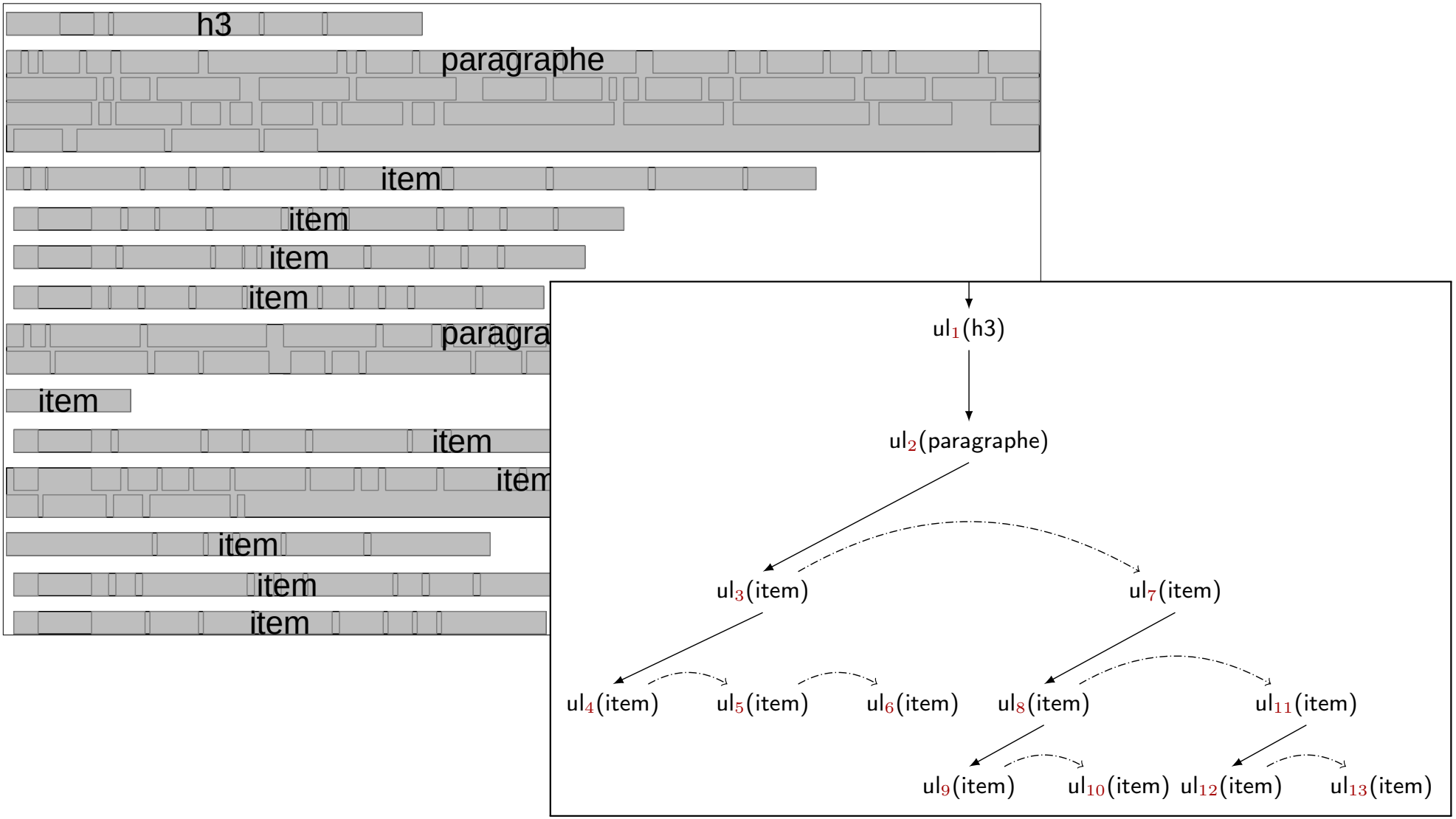
Configurations	LING	GEOP	LING_GEOP
Traits d'états	79,97	81,09	74,58
+ Traits de transitions	**87,24**	**82,88**	**80,23**
baseline	32,33	44,51	37,33

Courbes d'apprentissage



(2) Construction de l'arbre de dépendances

Lier les unités par des subordinations et des coordinations



(2) Construction de l'arbre de dépendances

Parsing et classification

Analyseur LR(1) avec deux méthodes pour la transposition des états du parseurs en actions :

- **Grammaire hors-contexte**
  - Série de règles simples déterminant les dépendances :
    $\text{ul}_i(\text{h1})\ \text{S}\left(\text{* ul}_j(\text{h2})\right)$
    ...
    $\text{ul}_i(\text{paragraphe})\ \text{S}\left(\text{* ul}_j(\text{item})\right)$
    ...
- **Classifieur statistique**
  - Traits : comparables à la tâche d'étiquetage
  - Classifieur : Régression Logistique Multinomiale

$$p(y|\mathbf{x}) = \frac{\exp(\theta_y^\mathsf{T}\mathbf{x})}{\sum_{c=1}^{|Y|}\exp(\theta_c^\mathsf{T}\mathbf{x})}$$

(2) Construction de l'arbre de dépendances

Évaluation (exactitude)

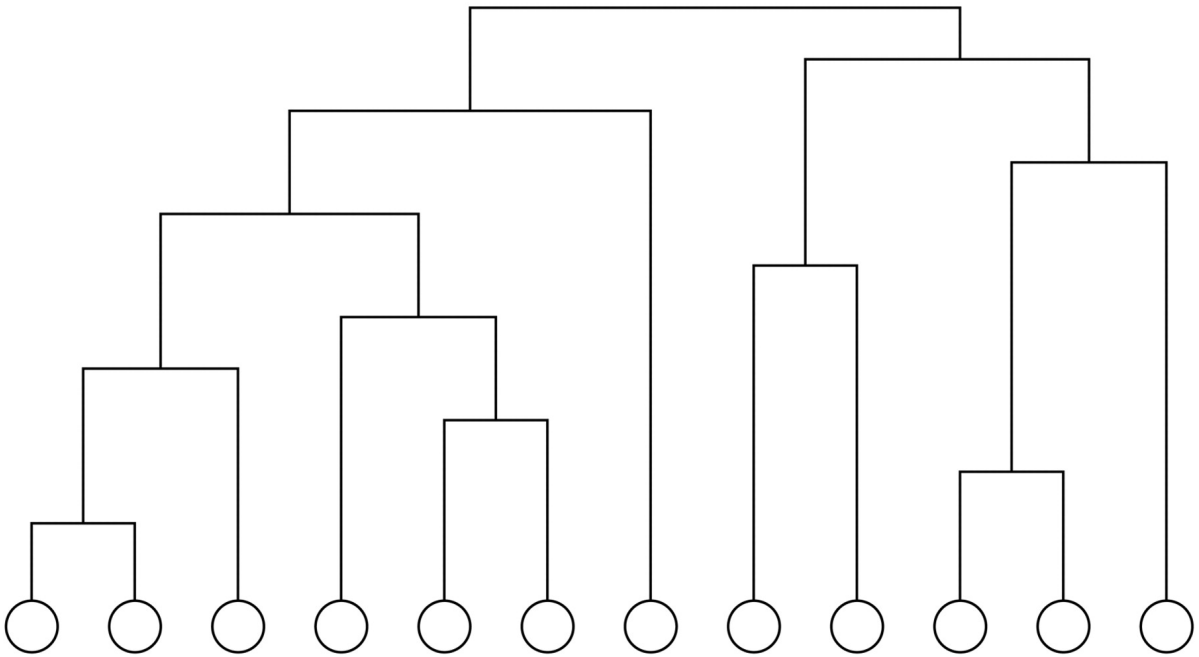
Méthodes	LING	GEOP	LING_GEOP
Grammaire	**96,54**	98,30	97,08
Classifieur statistique	96,41	**98,45**	**97,23**
Baseline	40,21	41,03	39,79

- **Grammaire** :
  - LING présente une structuration complexe
  - Les dépendances entre unités suivent majoritairement la grammaire
- **Classifieur statistique**
  - Asymétrie des données $\Rightarrow$ apprentissage de la grammaire
  - Amélioration modeste sur les cas difficiles : $\sim 15\%$

Structure de document

Perspectives

- Apprentissage sur un plus grand nombre de données
  - Acquisition non-supervisée de traits
- Apprentissage de classes d'équivalences visuelles
  - Premières expériences avec du clustering hiérarchique



Partie II

Extraction de Relations

Plan

Structure énumérative d'intérêt

Structure énumérative d'intérêt

Nous ciblons les SE présentant des propriétés rhétoriques et visuelles distinctes.

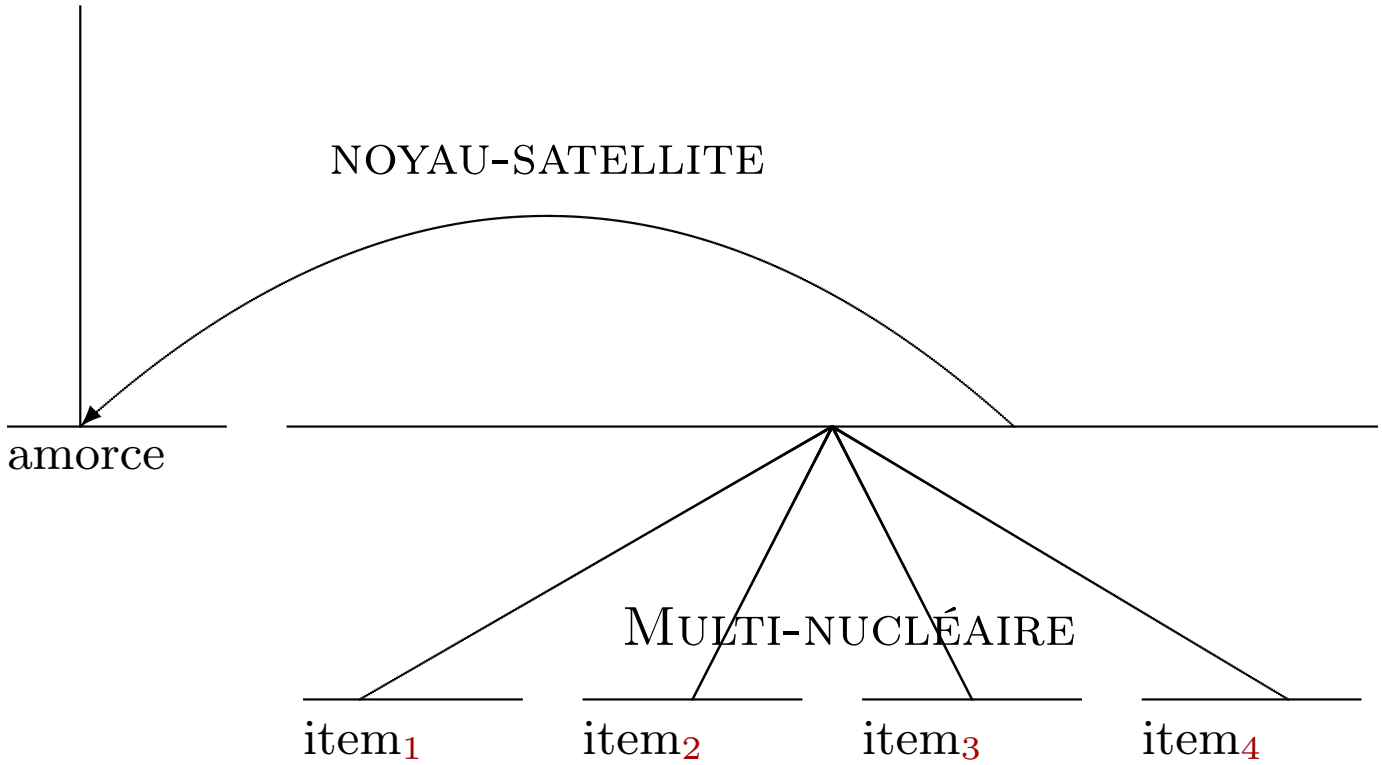
- Elles sont paradigmatiques
- Elles sont marquées typographiquement et dispostionnellement

Les formes de communication non parlées sont :

- le langage écrit
- le langage des signes
- le langage sifflé
- le langage du corps

Structure énumérative d'intérêt

Représentation rhétorique (RST) d'une SE d'intérêt



Paradigmatique (Luc, 2000)

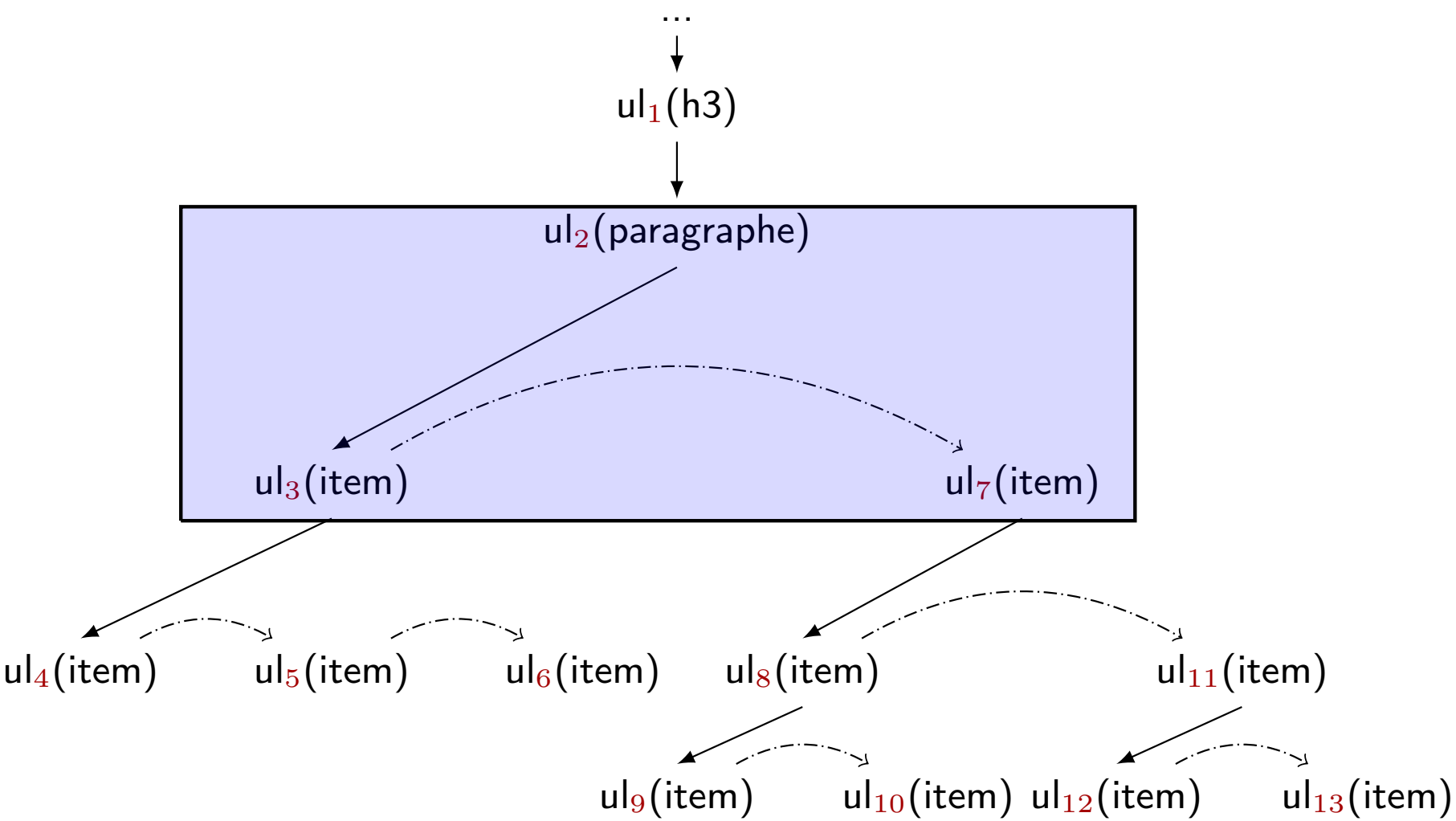
Items équivalents du point de vue fonctionnel

Absence de dépendances syntaxiques ou rhétoriques entre les items

Structure énumérative d'intérêt

Identification des structures énumératives d'intérêt

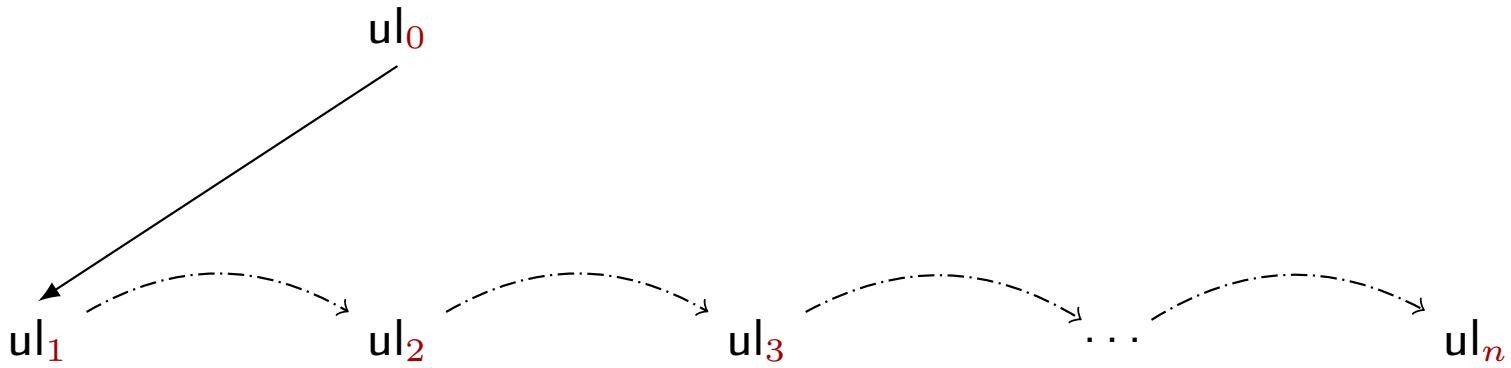
Indépendant des indices concrets de mise en forme



Structure énumérative d'intérêt

Filtrage par motifs

- n unités logiques coordonnées et étiquetées *item*
- la première unité logique est subordonnée à une unité logique père



Distribution de la relation sémantique

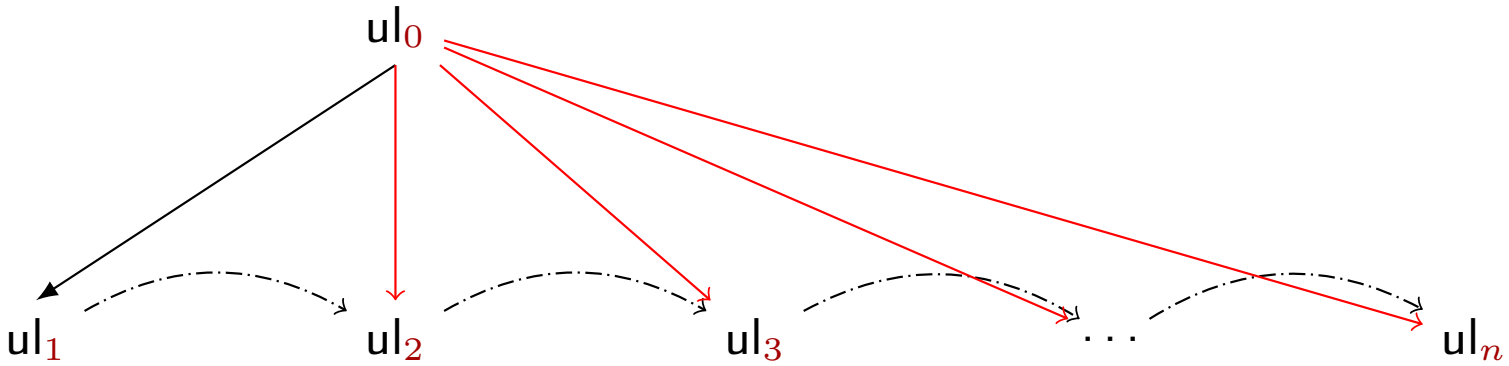
Quand cette structure porte une relation sémantique, généralement :

- ul$_0$ contient le classificateur (p. ex. hyperonyme)
- ul$_i$ $(1 \leq i \leq n)$ contiennent les entités textuelles co-énumérées (p. ex. hyponymes)

Structure énumérative d'intérêt

Filtrage par motifs

- n unités logiques coordonnées et étiquetées *item*
- la première unité logique est subordonnée à une unité logique père



Distribution de la relation sémantique

Quand cette structure porte une relation sémantique, généralement :

- ul$_0$ contient le classificateur (p. ex. hyperonyme)
- ul$_i$ $(1 \leq i \leq n)$ contiennent les entités textuelles co-énumérées (p. ex. hyponymes)

Structure énumérative d'intérêt

SE d'intérêt avec une relation d'hyperonymie

Liste des principaux estuaires de France

- Estuaire de la Garonne appelé aussi estuaire de la Gironde ;
- Estuaire de la Seine ;
- Estuaire de la Loire, partie aval de la Basse-Loire correspondant à l'embouchure de la Loire ;
- Estuaire de la Rance : voir aussi l'Usine marémotrice de la Rance ;
- La série des « Estuaire picards » à la configuration géomorphologique particulière (avec du sud au nord, les estuaires de la Somme, de la Canche, de l'Authie, de la Liane (artificialisé), de la Slack et du Wimmereux.

Structure énumérative d'intérêt

SE non paradigmatique

Est considéré comme « lecture savante », du point de vue fonctionnel, une pratique de lecture répondant aux critères suivants :

- c'est une lecture « qualifiée »,
- qui se développe sur le temps long de la recherche scientifique,
- dans un parcours forcément individualisé,
- où l'écriture se combine à la lecture, souvent dans une perspective de publications.

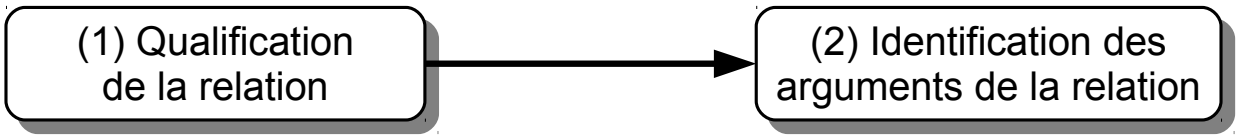
Plan

Extraction de relations dans les SE

Objectif

Identifier automatiquement les structures énumératives d'intérêt porteuses de relations sémantiques utiles à la construction de ressources.

Deux tâches



Expériences menées sur Wikipédia

- Pages des concepts de l'ontologie OntoTopo (ANR GeOnto)
- Campagne d'annotation :
  - 3 annotateurs et outils
  - 169 documents et 745 SE
  - accord : κ 0,54 pour l'hyperonymie
- Prétraitement avec Talismane (Urieli, 2013)

(1) Qualification de la relation sémantique

Classification (SAC, 2015)

Qualifier la nature de la relation sémantique portée par les structures énumératives verticales identifiées.

- Chaque SE est représentée par :

  catégories grammaticales, marqueurs lexicaux, saturation syntaxique, etc.

Évaluation pour l'hyperonymie

Deux classifieurs : Régression logistique et SVM à noyau gaussien

Stratégies	Précision	Rappel	F_1-score	Exactitude
Régression logistique	78,01	84,78	81,25	**75,34**
SVM	74,77	90,22	**81,77**	74,66
Baseline	63,01	100,0	77,31	63,01

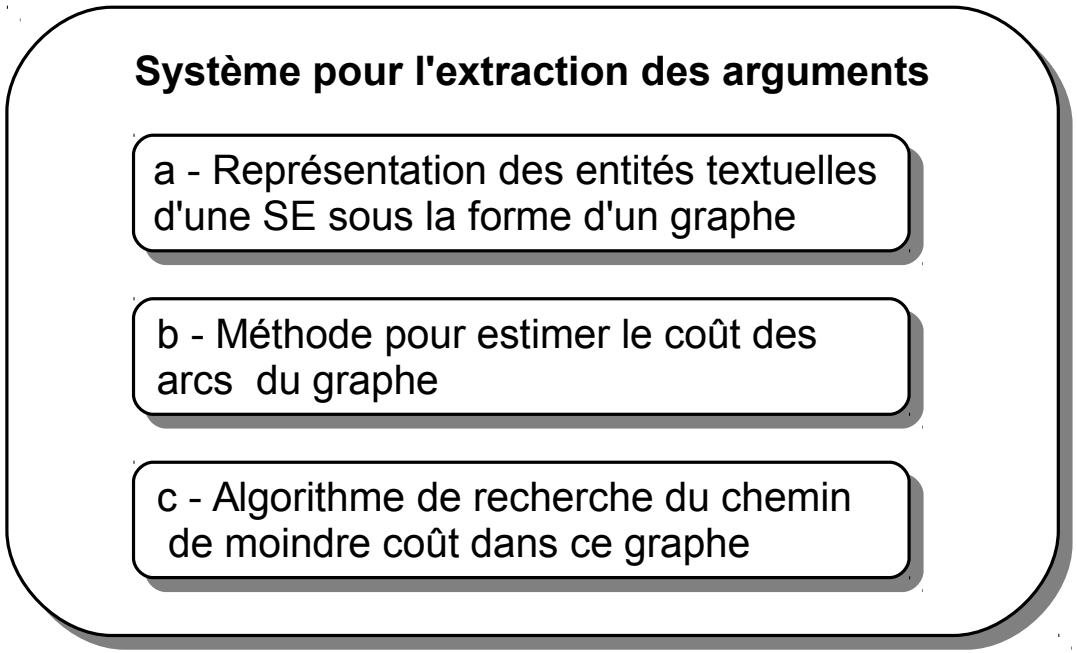
Voir (TALN, 2013) et (RIA, 2014) pour autres types de relations

(2) Identification des arguments de la relation

Prédiction structurée (*SEM, 2015)

Identification des entités textuelles impliquées dans les relations sémantiques au travers d'un système de prédiction structurée

- Hypothèse : les arguments apparaissent dans le chemin d'entités montrant la plus grande cohésion lexicale et dispositionnelle



(2) Identification des arguments de la relation

a - Représentation des SE sous la forme de graphes

Identification des entités textuelles

Extracteurs : ACABIT (Daille, 1996), YaTeA (Aubin et Hamon, 2006)

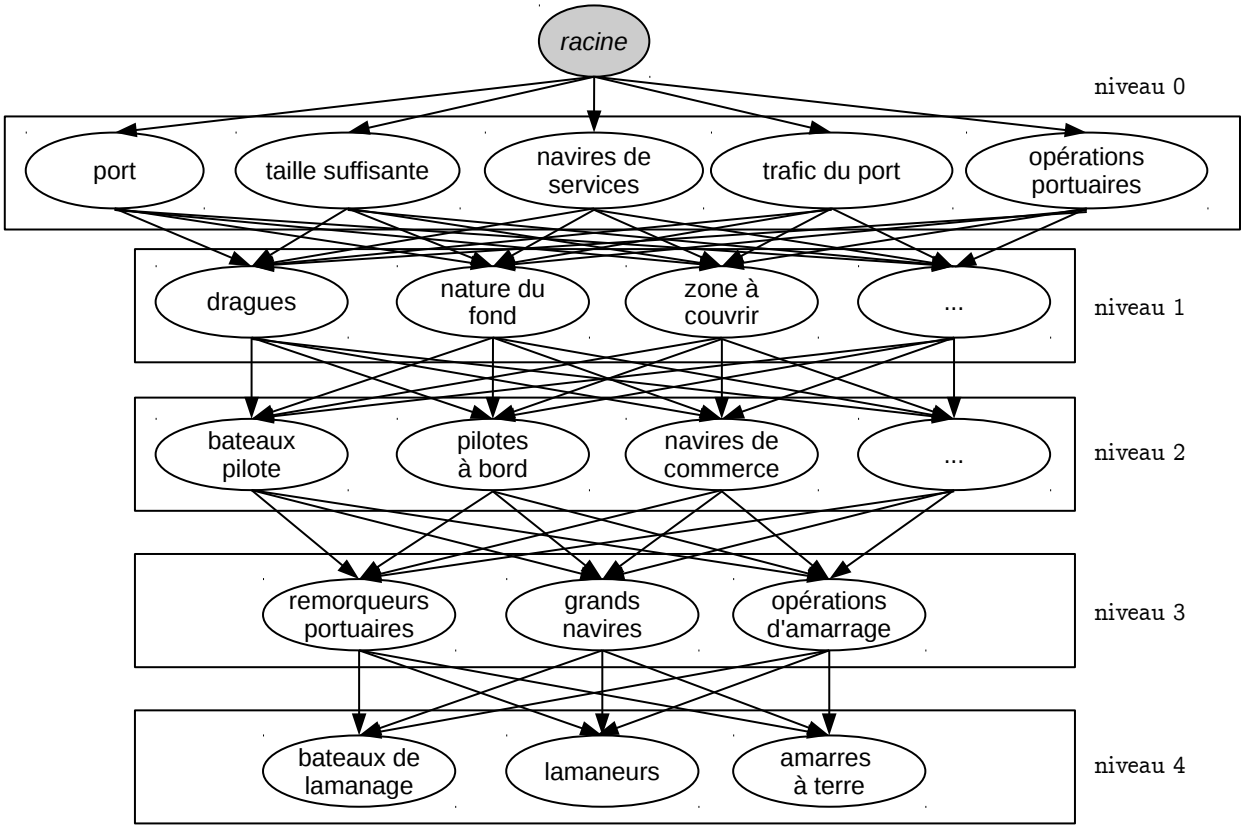
Dès qu'un port atteint une taille suffisante, un certain nombre de navires de services y sont basés ; ils ne font pas partie du trafic du port mais sont utilisés pour différentes opérations portuaires. On trouve ainsi :

- Les dragues, de différents types suivant la nature du fond et la zone à couvrir (à élinde traînante, à godets. . .) ; elles servent à maintenir une profondeur suffisante dans le port et les chenaux d'accès, malgré l'apport de sédiments dû aux rivières et courants.
- Les bateaux pilote servant à amener les pilotes à bord des navires de commerce arrivant au port. Sur les ports de moyenne importance, on trouve quelques pilotines opérant à partir du port ; sur les grands ports de commerce, on trouve parfois un grand navire dans la zone d'atterrissage hébergeant les pilotes, et duquel partent les pilotines.
- Les remorqueurs portuaires qui servent à aider les grands navires à manœuvrer durant les opérations d'amarrage et d'évitage.
- Les bateaux de lamanage utilisés par les lamaneurs pour porter les amarres à terre.

(2) Identification des arguments de la relation

a - Représentation des SE sous la forme de graphes

Transformation en un graphe orienté acyclique



Équivalences

- SE verticale = graphe
- Unité logique = niveau
- Entités textuelles = nœuds
- Liens possibles = arcs

(2) Identification des arguments de la relation

b - Méthode pour estimer le coût des arcs

Attribuer un coût aux arcs

$$\text{cost}(< T_i^j, T_{i+1}^k >) = 1 - p(y|T_i^j, T_{i+1}^k)$$

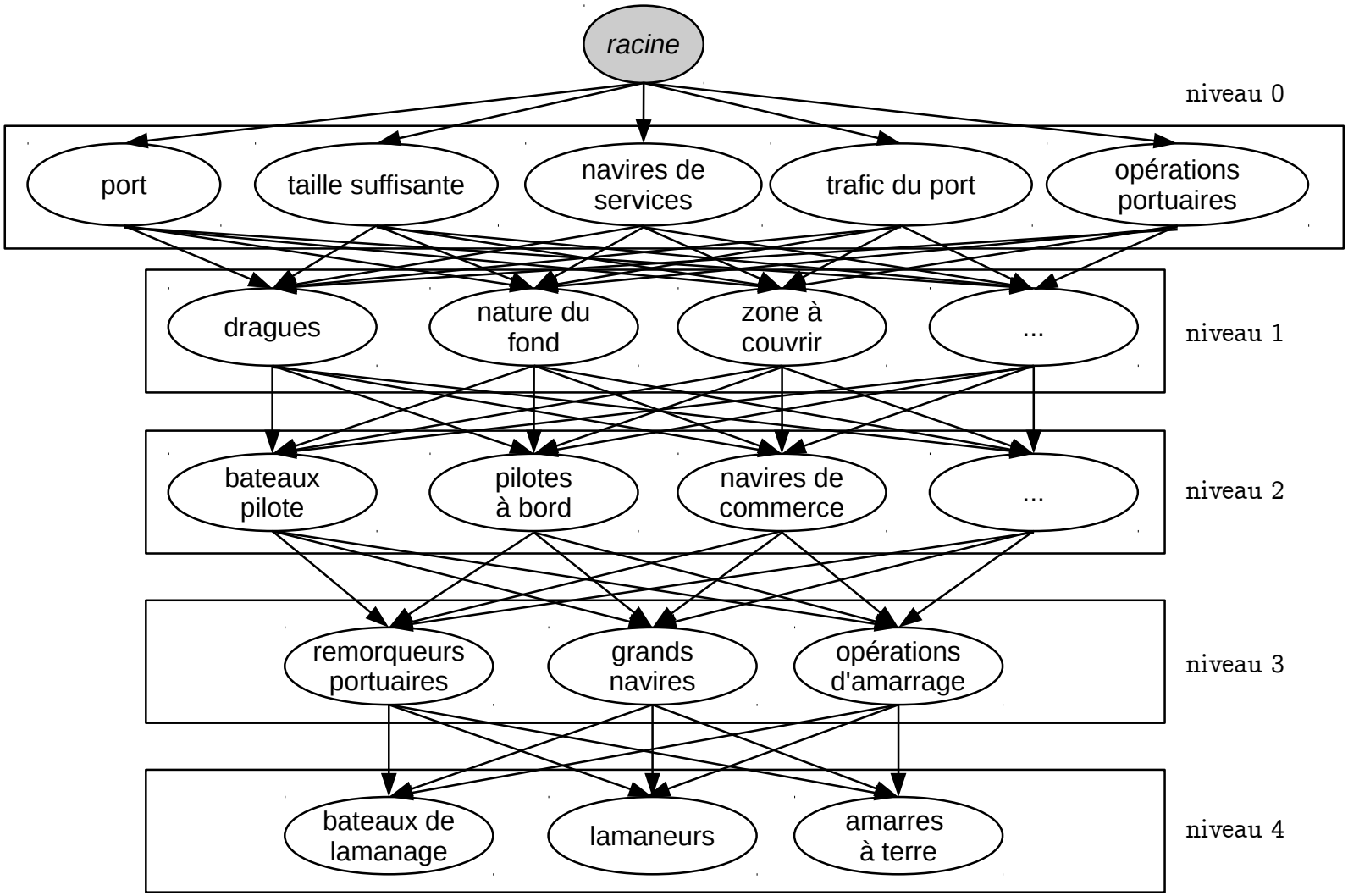
où T_i^j est la j-ème entité textuelle dans l'unité logique de niveau i et y représente l'événement où les deux entités textuelles soient impliquées dans la relation sémantique.

Estimation de la probabilité

- Deux régressions logistiques pour :
  1. paire hyperonyme - hyponyme
  2. paire hyponyme - hyponyme
- Traits : indices typographiques, dispositionnels, lexicaux et syntaxiques
  e.x. contexte morpho-syntaxique, position, catégories grammaticales, etc.

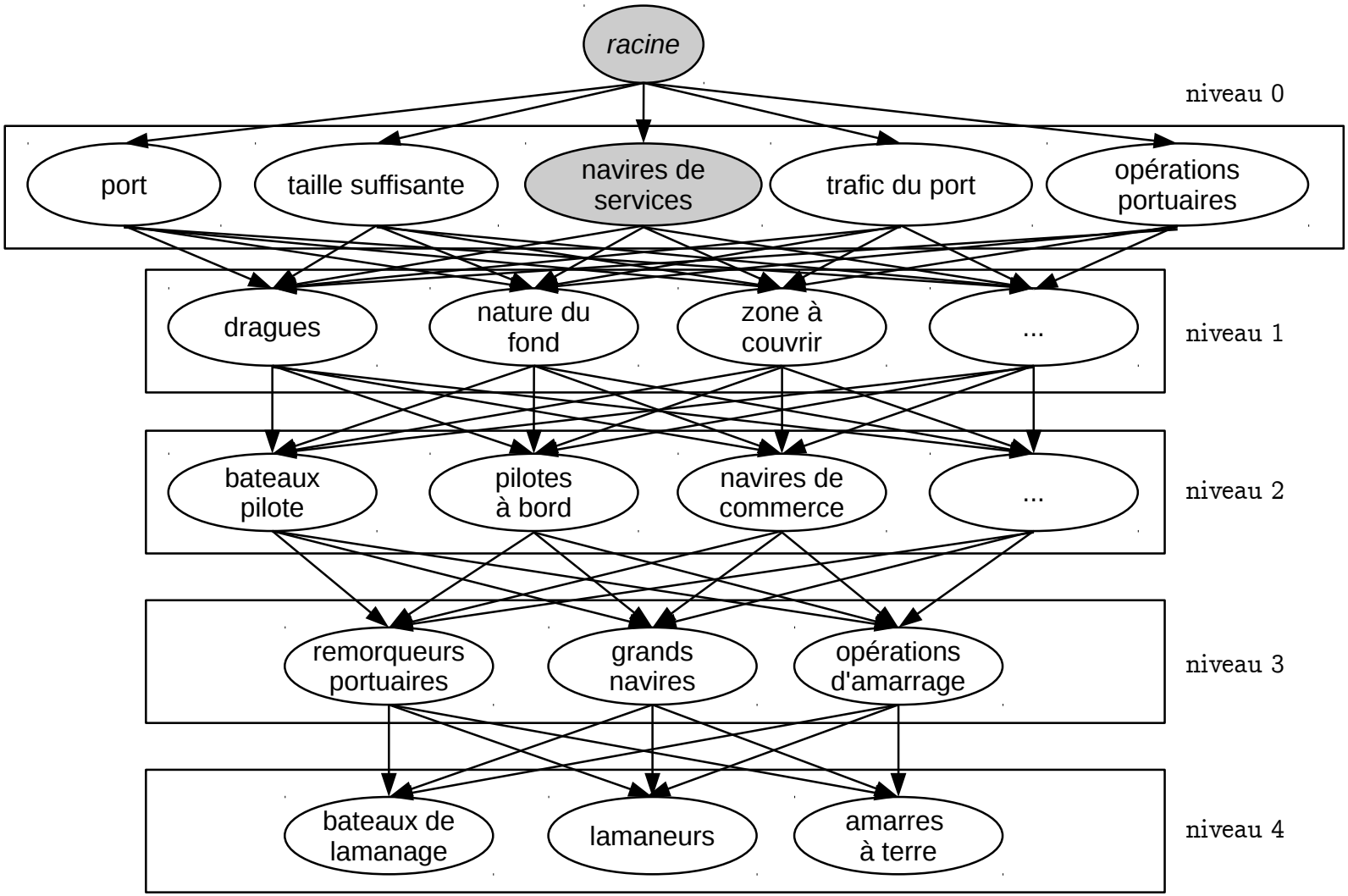
(2) Identification des arguments de la relation

c - Algorithme de recherche du chemin de moindre coût : A*



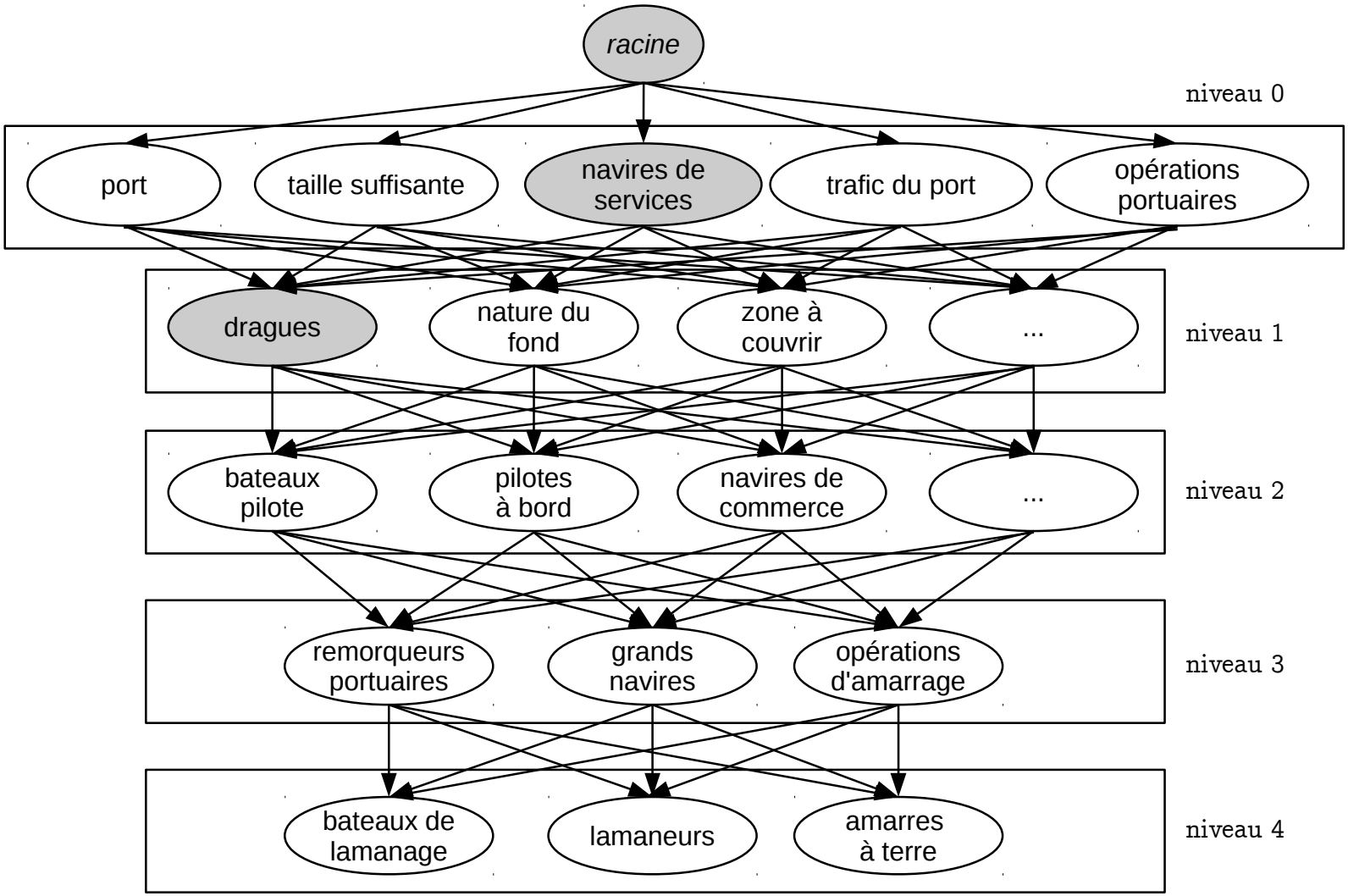
(2) Identification des arguments de la relation

c - Algorithme de recherche du chemin de moindre coût : A*



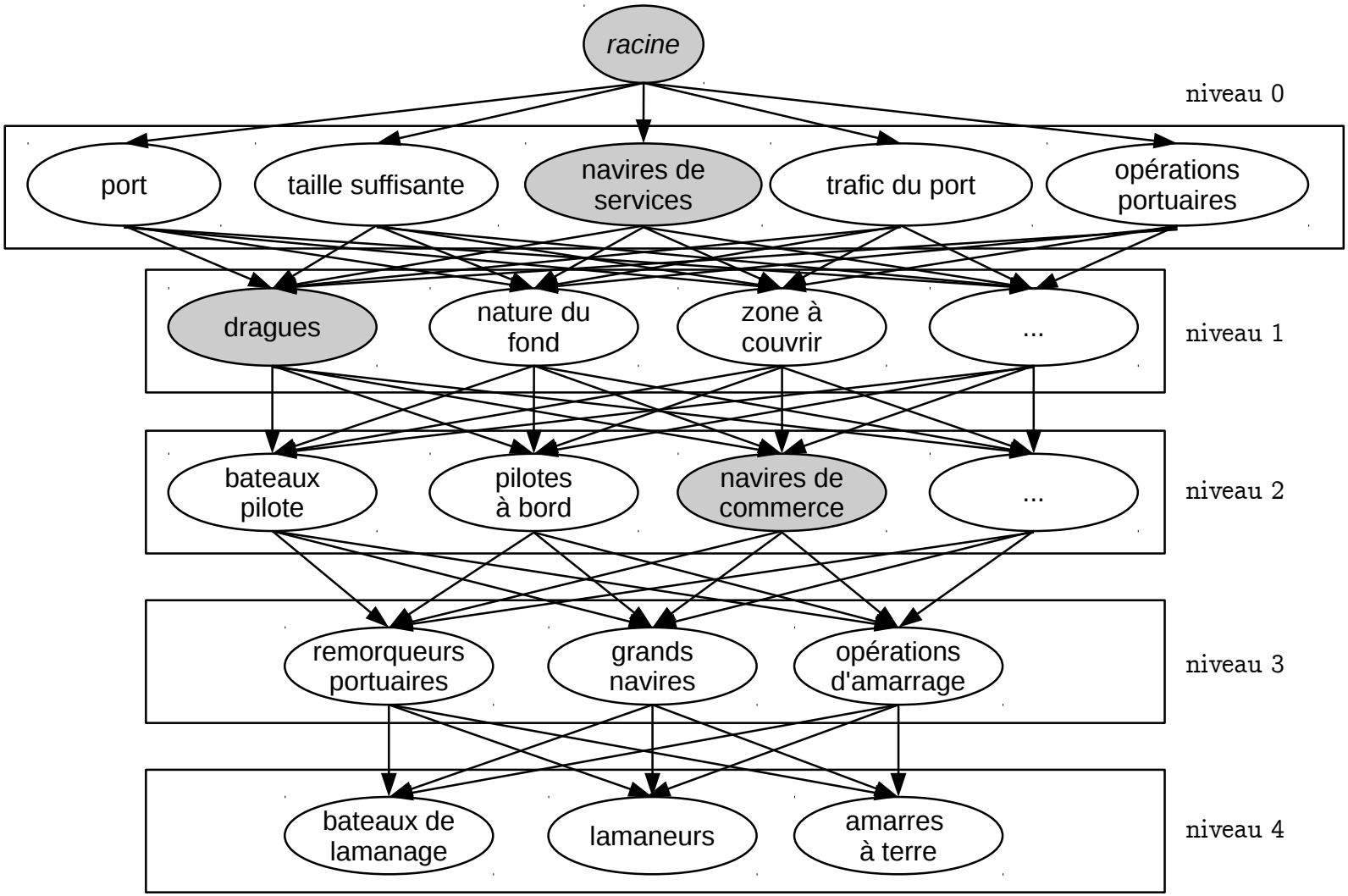
(2) Identification des arguments de la relation

c - Algorithme de recherche du chemin de moindre coût : A*



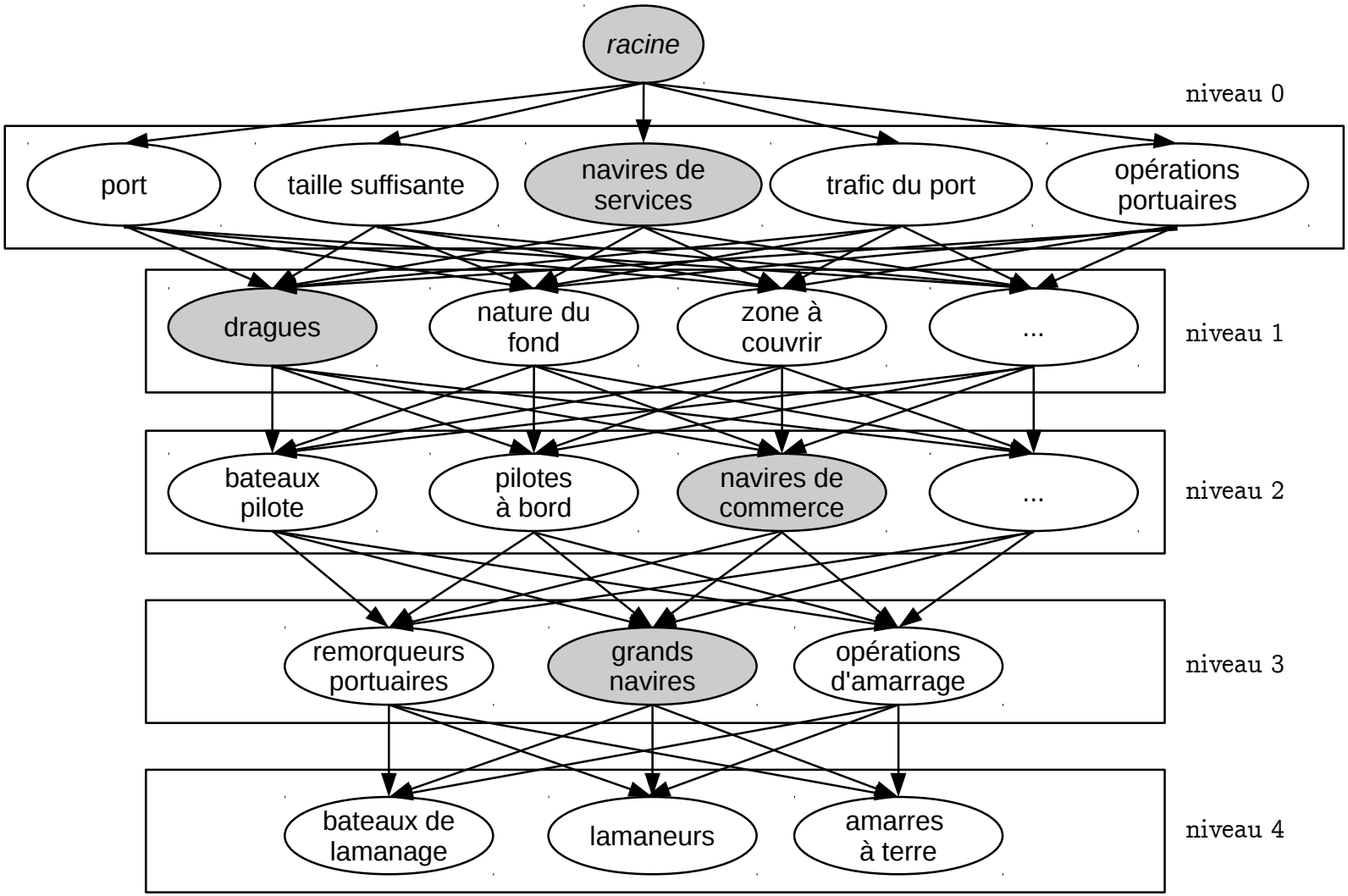
(2) Identification des arguments de la relation

c - Algorithme de recherche du chemin de moindre coût : A*



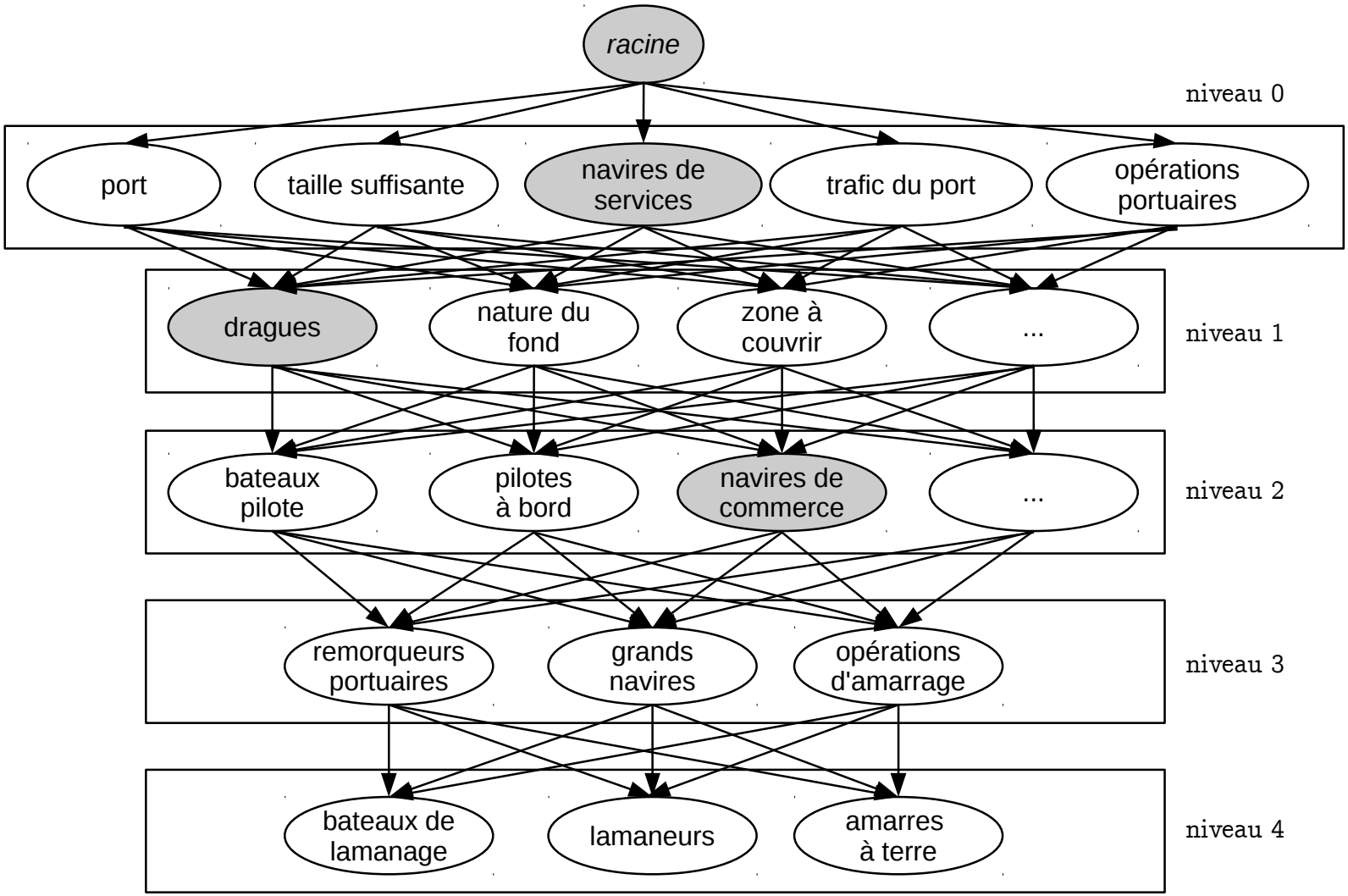
(2) Identification des arguments de la relation

c - Algorithme de recherche du chemin de moindre coût : A*



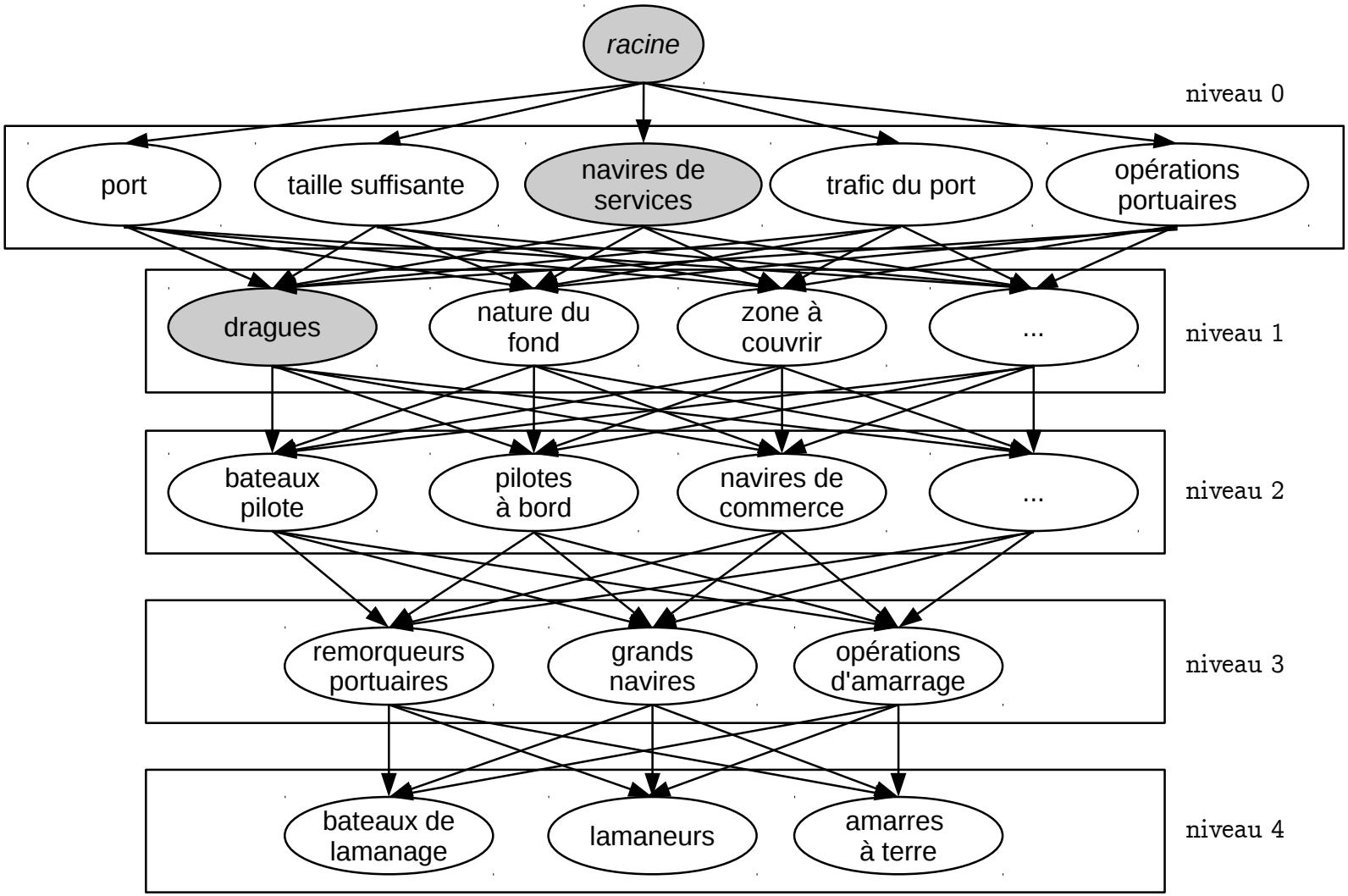
(2) Identification des arguments de la relation

c - Algorithme de recherche du chemin de moindre coût : A*



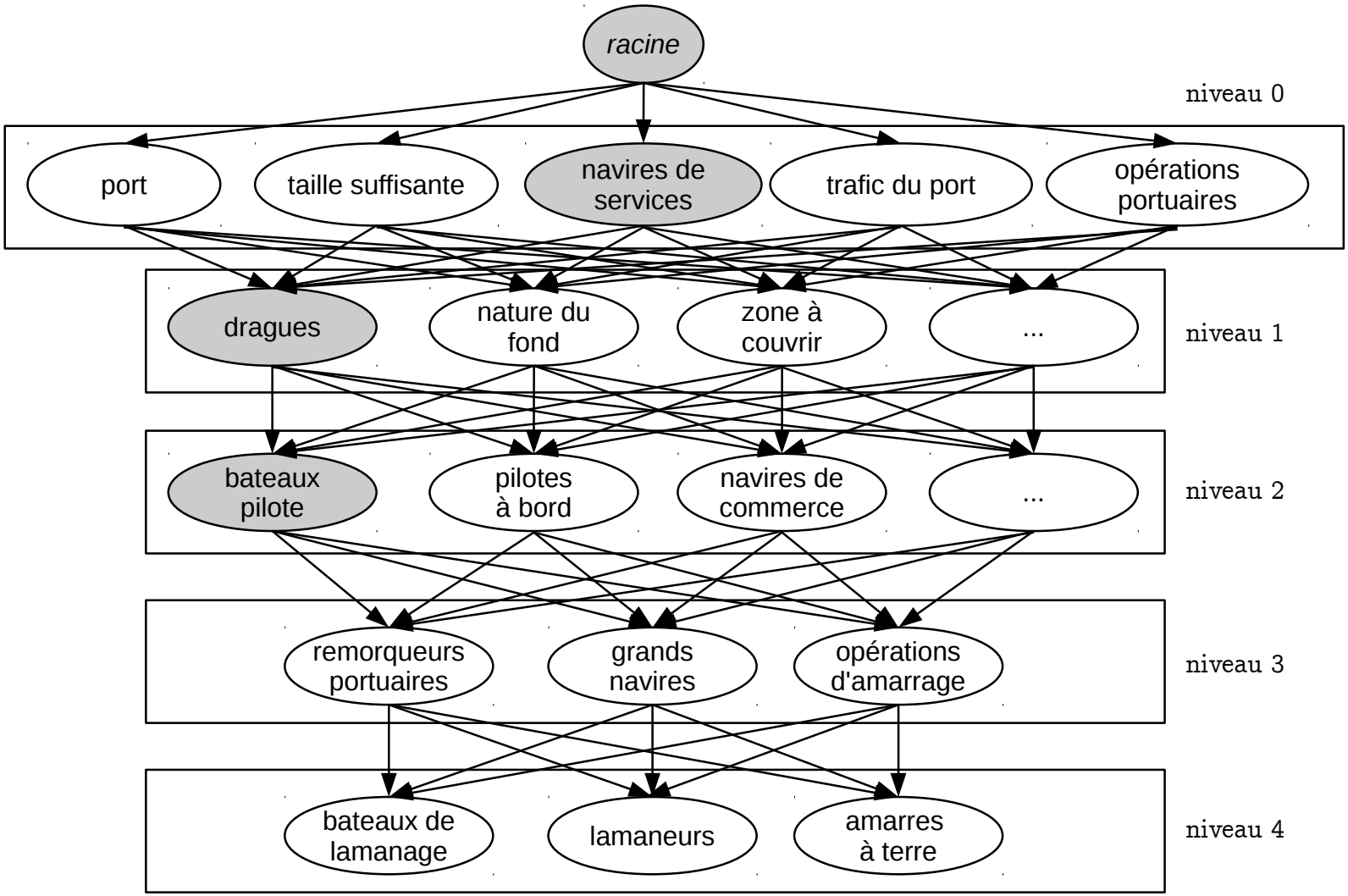
(2) Identification des arguments de la relation

c - Algorithme de recherche du chemin de moindre coût : A*



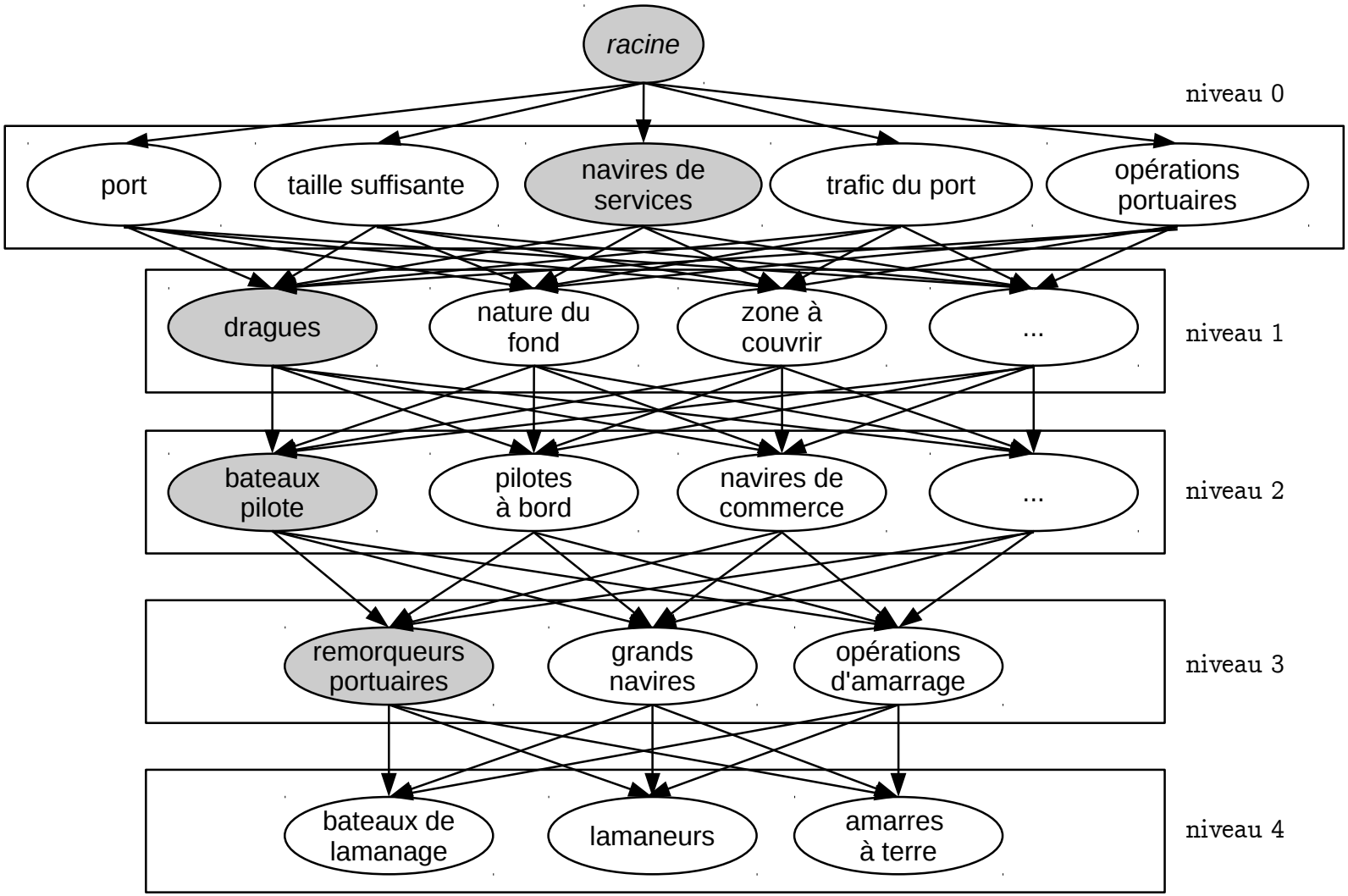
(2) Identification des arguments de la relation

c - Algorithme de recherche du chemin de moindre coût : A*



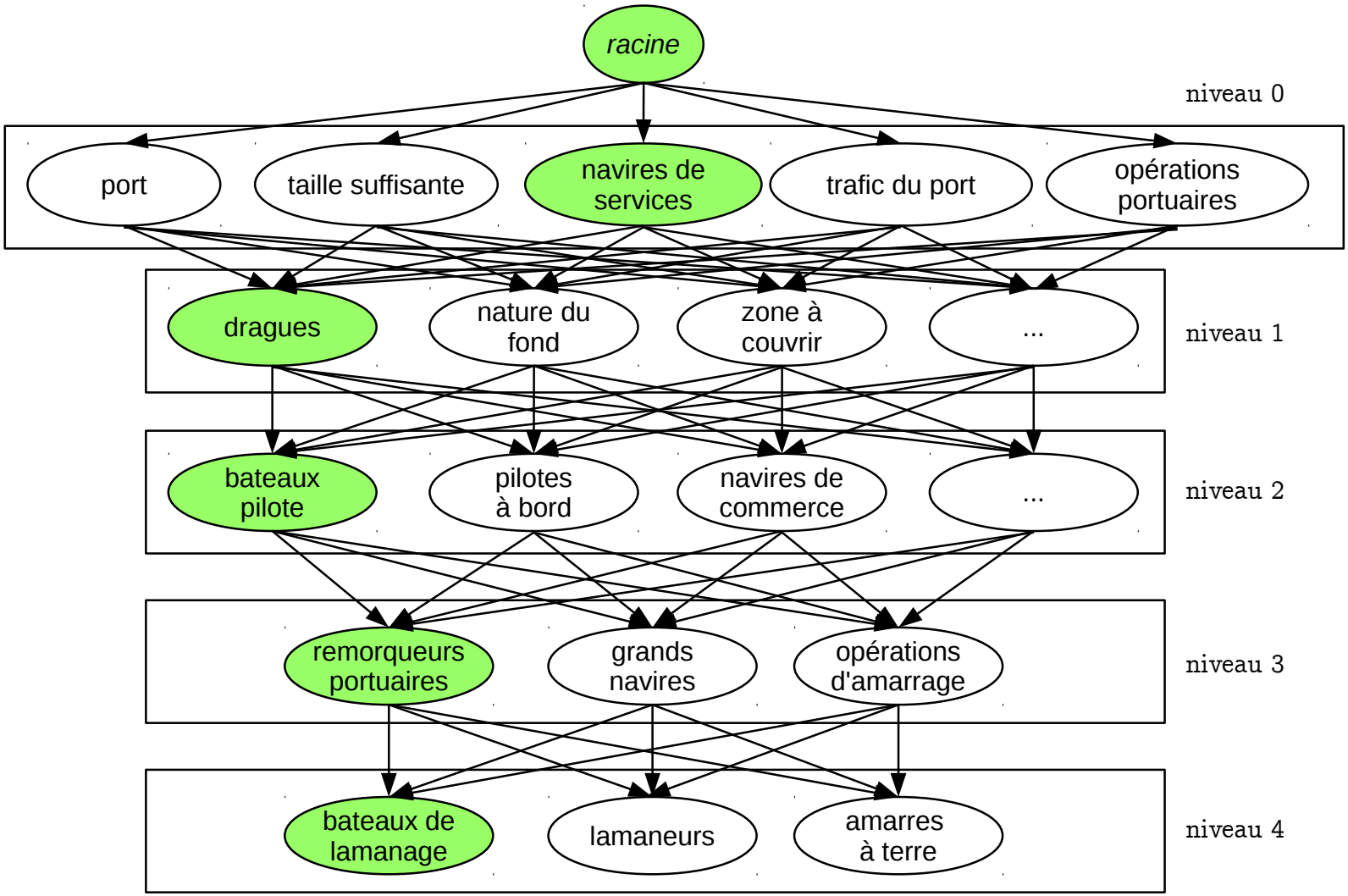
(2) Identification des arguments de la relation

c - Algorithme de recherche du chemin de moindre coût : A*



(2) Identification des arguments de la relation

c - Algorithme de recherche du chemin de moindre coût : A*



(2) Identification des arguments de la relation

Évaluation

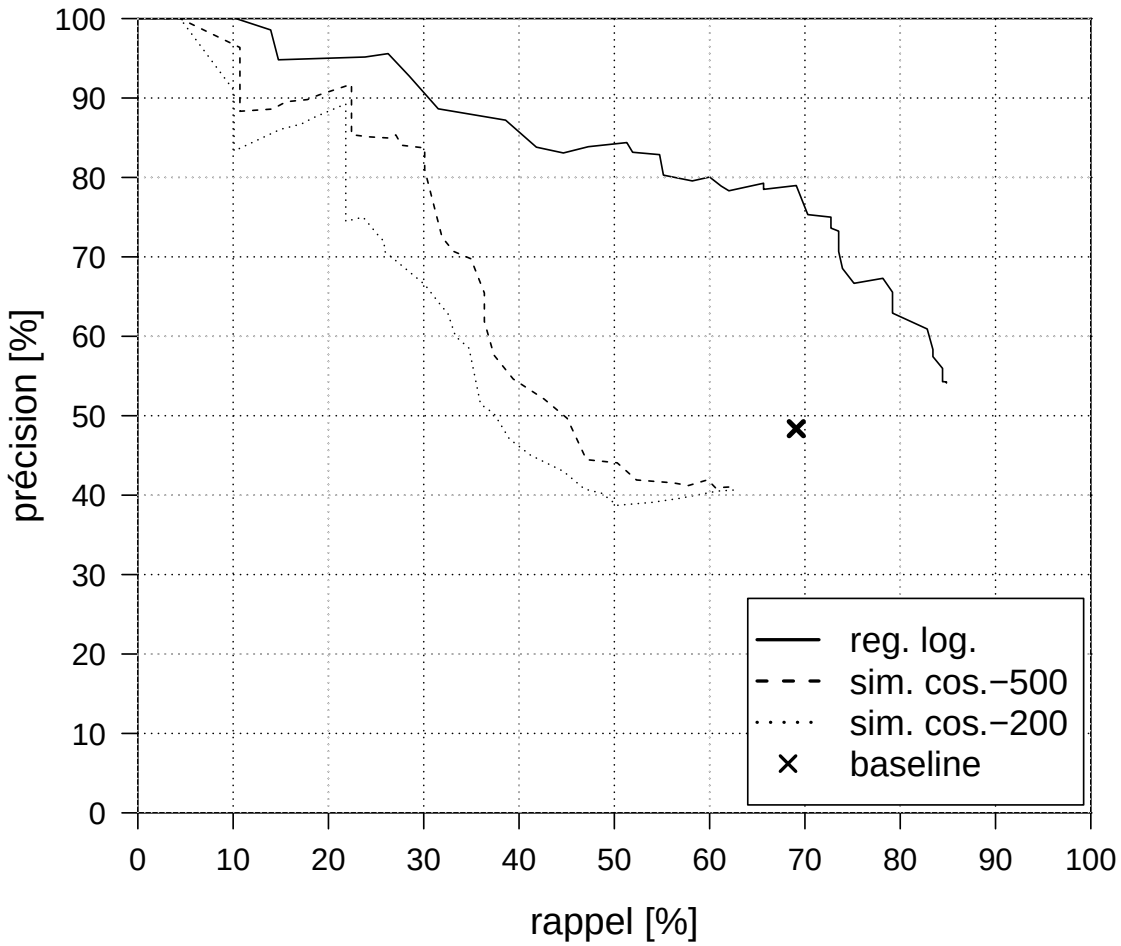
- Alternative 1 : Baseline dispositionnelle
- Alternative 2 : Modèles distributionnels (Word2Vec + FrWac)

Stratégies	Précision	Rappel	F_1-score
Baseline dispositionnnelle	48,37	69,09	56,91
Similarité cosinus (dim, 500)	**83,71**	30,10	44,28
Similarité cosinus (dim, 200)	66,52	30,10	41,45
Régression logistique	78,98	**69,09**	**73,71**

- Baseline :
  - Les informations importantes sont généralement mises en avant
- Modèles distributionnels :
  - La cohésion lexicale est un bon indice
- Régression logistique :
  - Obtient les meilleurs scores

(2) Identification des arguments de la relation

Courbes obtenues en faisant varier le score de confiance



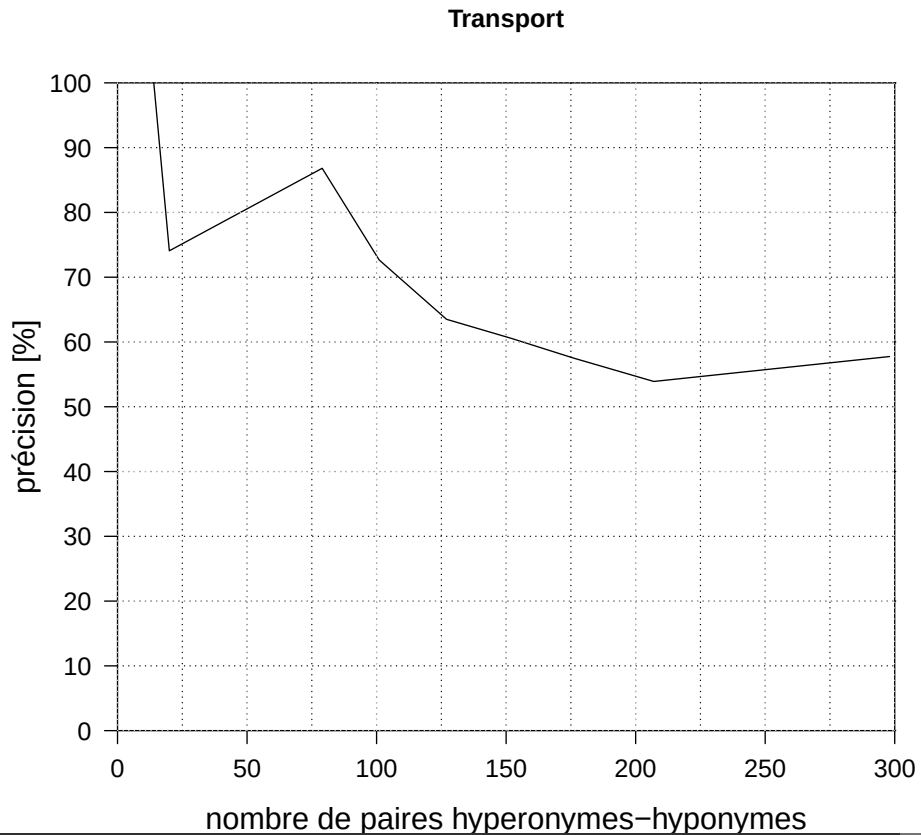
Plan

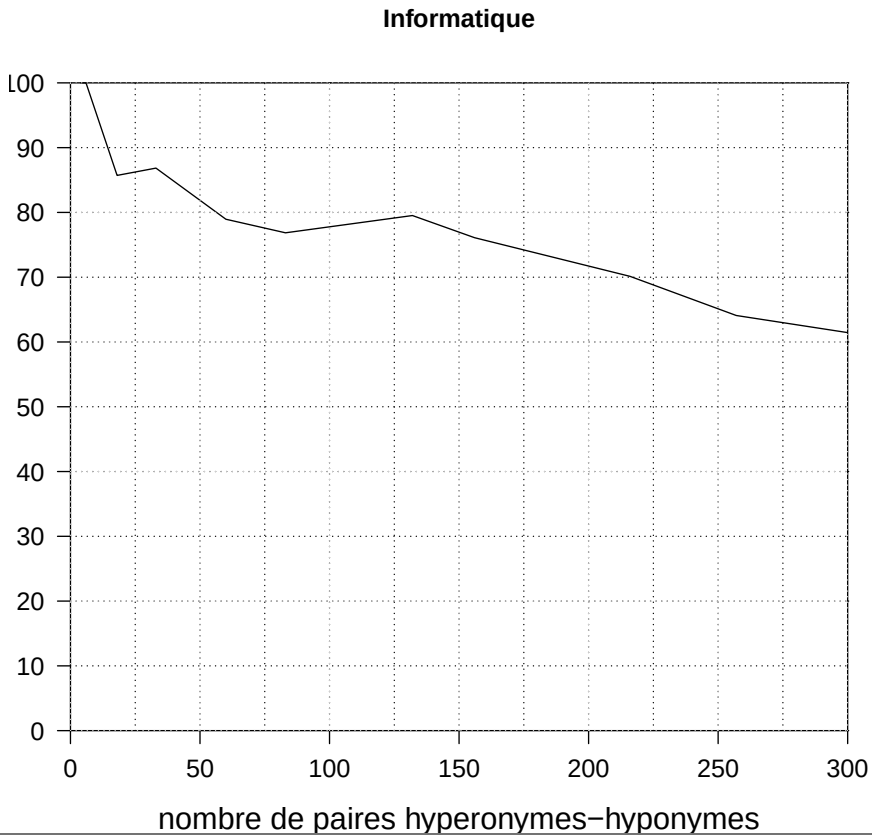
Évaluation de l'ensemble du système

Évaluation quantitative

Évaluation sur de nouvelles données

- Données de Wikipédia : 2 domaines x 400 documents
- Ensemble de la chaîne de traitement + tri par score de confiance
- Évaluation manuelle des 500 premières paires retournées





Évaluation de l'ensemble du système

Évaluation qualitative

SE imbriquées

Les entités textuelles directement co-énumérées caractérisent l'hyperonyme (« differentia » de (Bush, 2003))

- transmission sans fil
  - Courte distance
    - Bluetooth
  - Moyenne distance
    - Wi-Fi, 802.11
    - MANET
  - Longue distance
    - MMDS
    - SMDS
    - Transmission de données sur téléphone cellulaire
    - Réseaux de téléavertissement

Évaluation de l'ensemble du système

Évaluation qualitative

Confusions avec l'holonymie

Généralement présente dans les SE (Gala, 2003)

- Absence de marqueurs lexicaux
- Nécessité d'informations d'arrière-plan

- Direction générale de l'aviation civile : supervise le bureau des enquêtes chargé des investigations sur les accidents et incidents aériens graves survenant sur le territoire national
  - Service des affaires générales
  - Bureau des enquêtes
  - Direction des études et de l'exploitation du transport aérien
  - Direction du personnel aéronautique et du matériel volant
  - Direction de la navigation aérienne

Évaluation de l'ensemble du système

Évaluation qualitative

Phénomènes linguistiques

Imbrication de structures + expression d'une exclusion

- membre supérieur : en dehors des traumatismes bénins, on retrouve :
  - fracture de la palette humérale
  - rupture de la coiffe des rotateurs
  - fracture de la clavicule
  - fractures du poignet, le plus souvent du scaphoïde

Plan

Conclusion

Apports de la mise en forme dans l'extraction de relations

- Repérage des structures textuelles : ++
- Qualification de la relation sémantique : +
- Identification des arguments : ++

Travaux actuels

- Apprentissage de classes d'équivalences visuelles
- Représentation vectorielle de la mise en forme

Conclusion

Merci pour votre attention

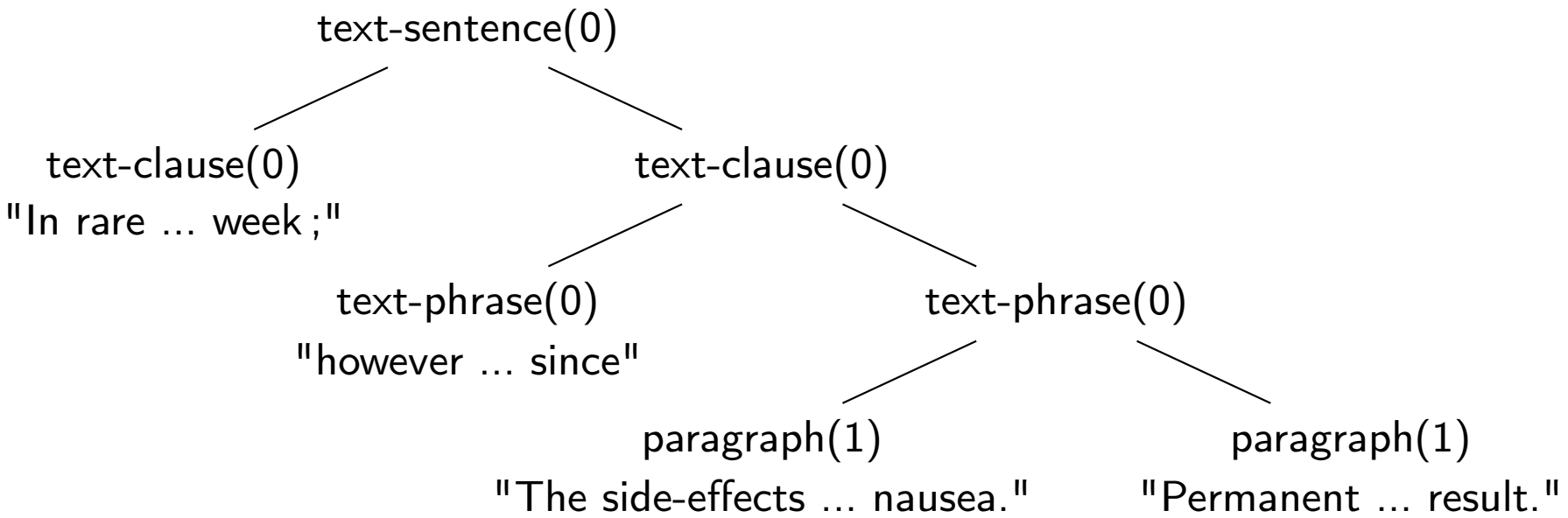
Annexes 1 : Structure de document

Annexes pour la structure de document :

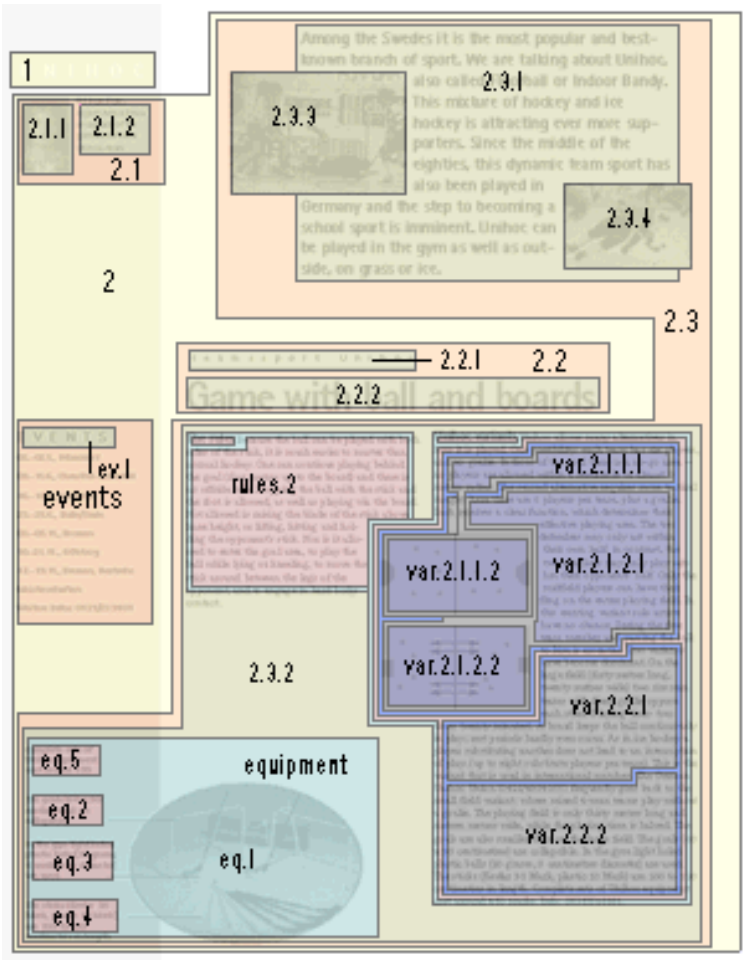
Modèle de Power *et al.* (2003)

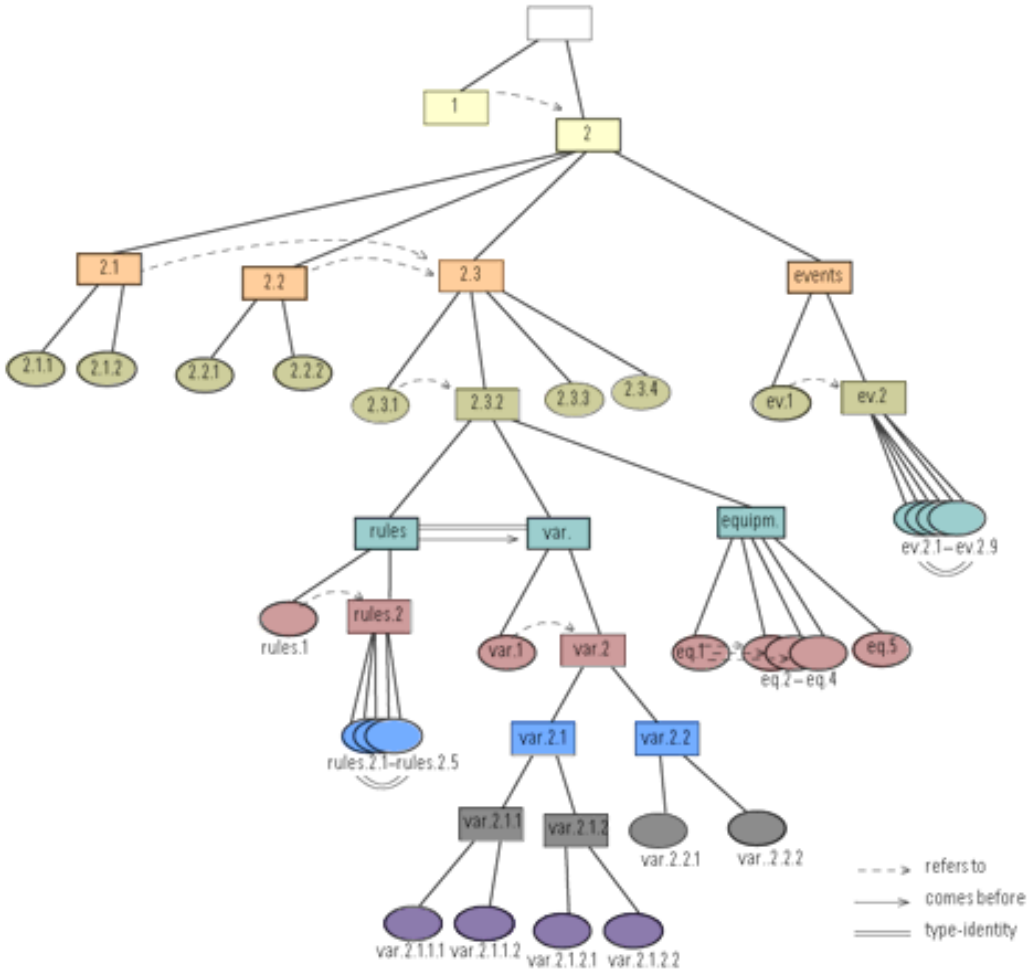
In rare cases the treatment can be prolonged for another week ; however, this is risky since

- The side-effects are likely to get worse. Some patients have reported severe headache and nausea.
- Permanent damage to the liver might result.

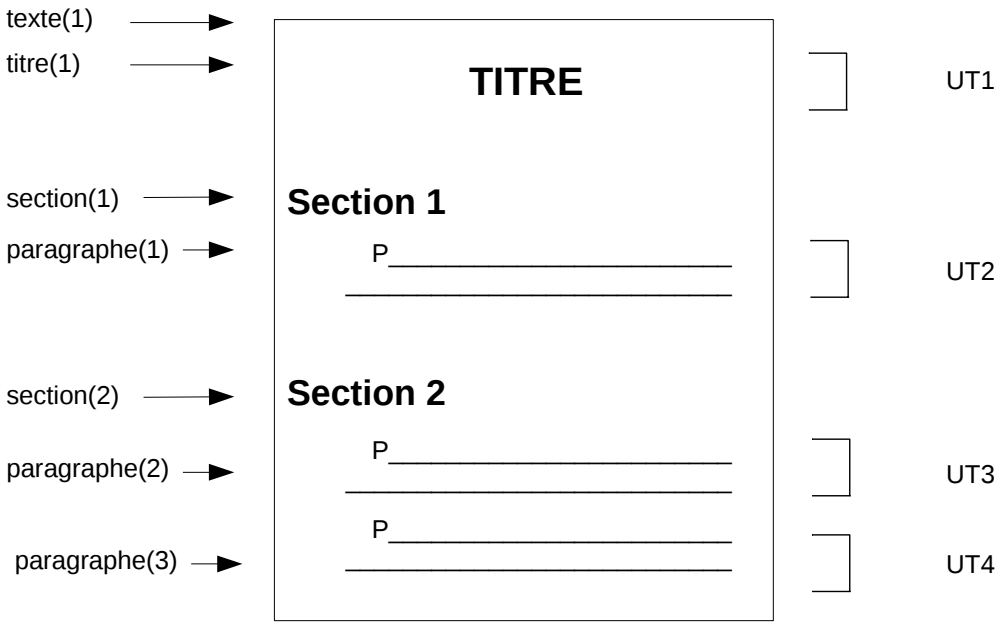


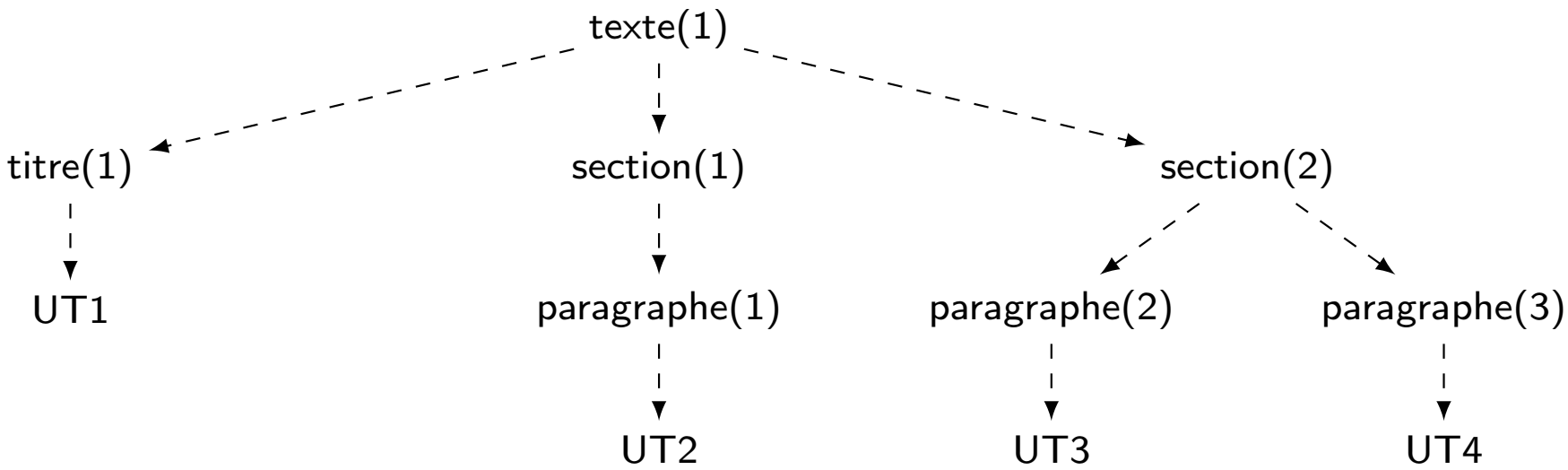
Modèle de Bateman *et al.* (2001)



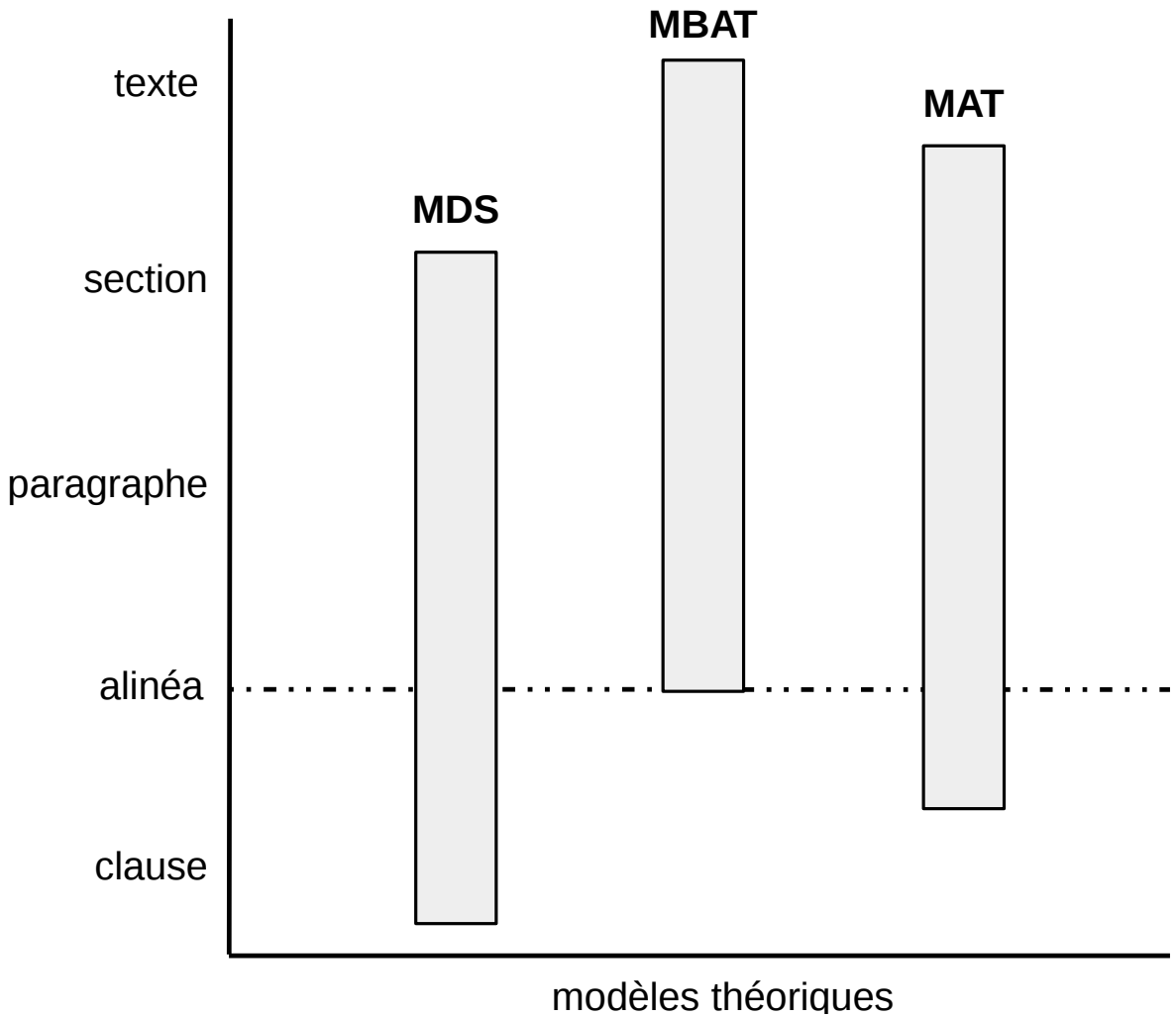


Modèle de Virbel (1989)





Comparaison des modèles TAL



Contraintes supplémentaires

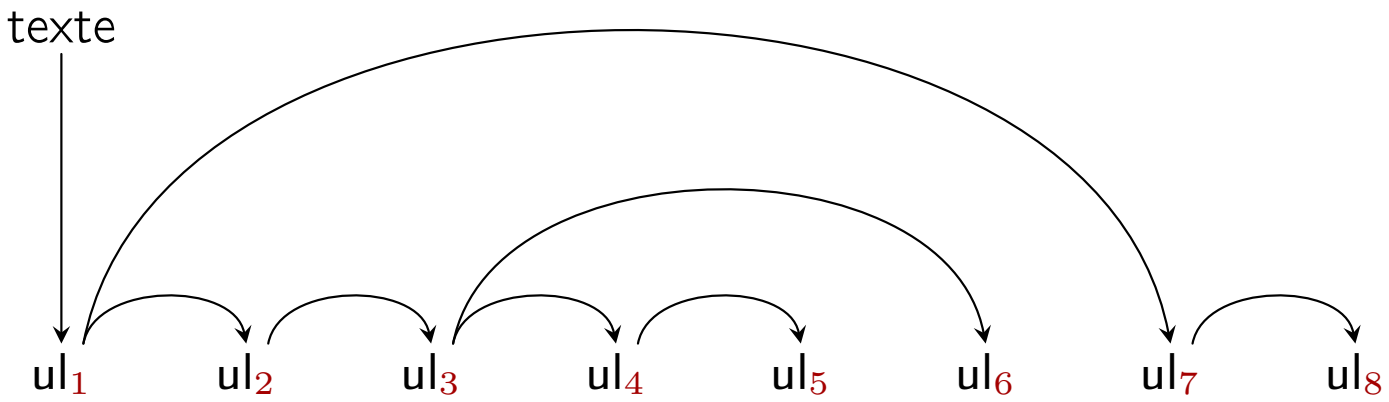


Figure – Arbre de dépendances projectif

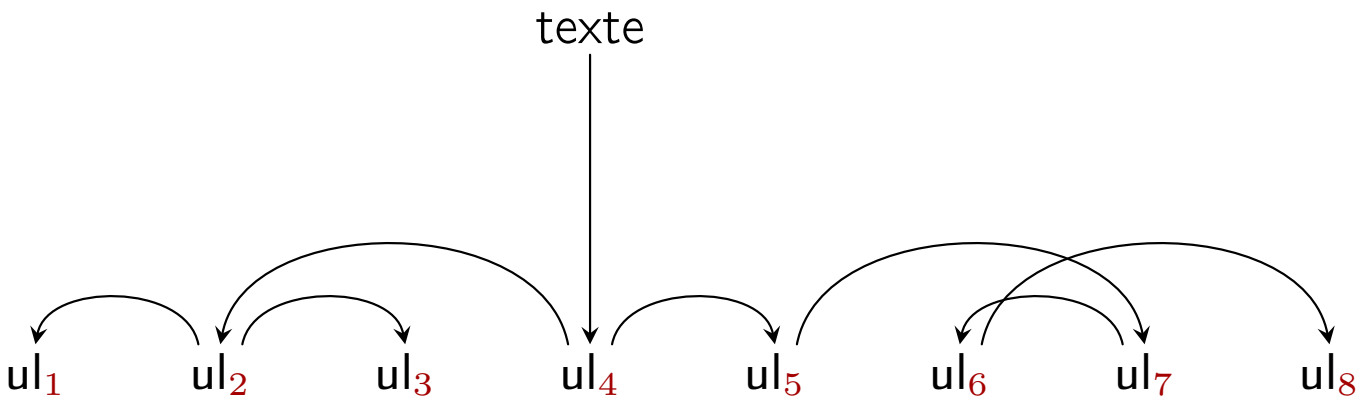


Figure – Arbre de dépendances non-projectif

Corpus LING_GEOP : (1) annotation visuelle

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 XML des propriétés visuelles :

- Caractérisation dispositionnelle en pixels : (x_1,y_1) (x_2,y_2)
- Caractérisation typographique pour les mots : *fonte*, *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>
```

Corpus LING_GEOP : (2) annotation de la structure logique de surface

Étiquetage avec étiquettes logiques :

1. Réutilisation des étiquettes présentes dans ANNODIS
   Réalisée avec un algorithme de similarité textuelle (Myers, 1986)
2. Ajout manuel des labels étiquettes pas traitées 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

Corpus LING_GEOP : (2) annotation de la structure logique de surface

étiquettes	LING (25 doc.) Nombre	Moyenne	GEOP 21 (doc.) Nombre	Moyenne	Total	Couv. %
titre	27	1,08	28	1,33	55	0,84
byline	53	2,12	96	4,57	149	2,29
réf. biblio.	1173	**46,92**	25	1,19	1198	18,39
h1	157	6,28	101	4,81	258	3,96
h2	108	4,32	78	3,71	186	2,85
h3	40	1,6	65	3,10	105	1,61
paragraphe	1241	49,64	1189	56,62	2430	**37,30**
item	380	**15,2**	72	3,43	452	6,94
citation	123	**4,92**	1	0,05	124	1,90
en-tête	45	1,8	171	**8,14**	216	3,32
pied de page	16	0,64	257	**12,24**	273	4,19
note de bas de page	7	0,28	370	**17,62**	377	5,79
note de fin	387	**15,48**	28	1,33	415	6,37
autres	82	3,28	195	**9,29**	277	4,25
Total	3839	153.56	2676	127.43	6515	100

Corpus LING_GEOP : (3) annotation de la structure logique profonde

Construction de l'arbre de dépendances :

1. Génération des relations entre unités logiques
   Réalisé avec une grammaire hors-contexte
2. Ajout manuel des structures multi-échelles de ANNODIS
   Uniquement celles dont la granularité est supérieure au bloc visuel

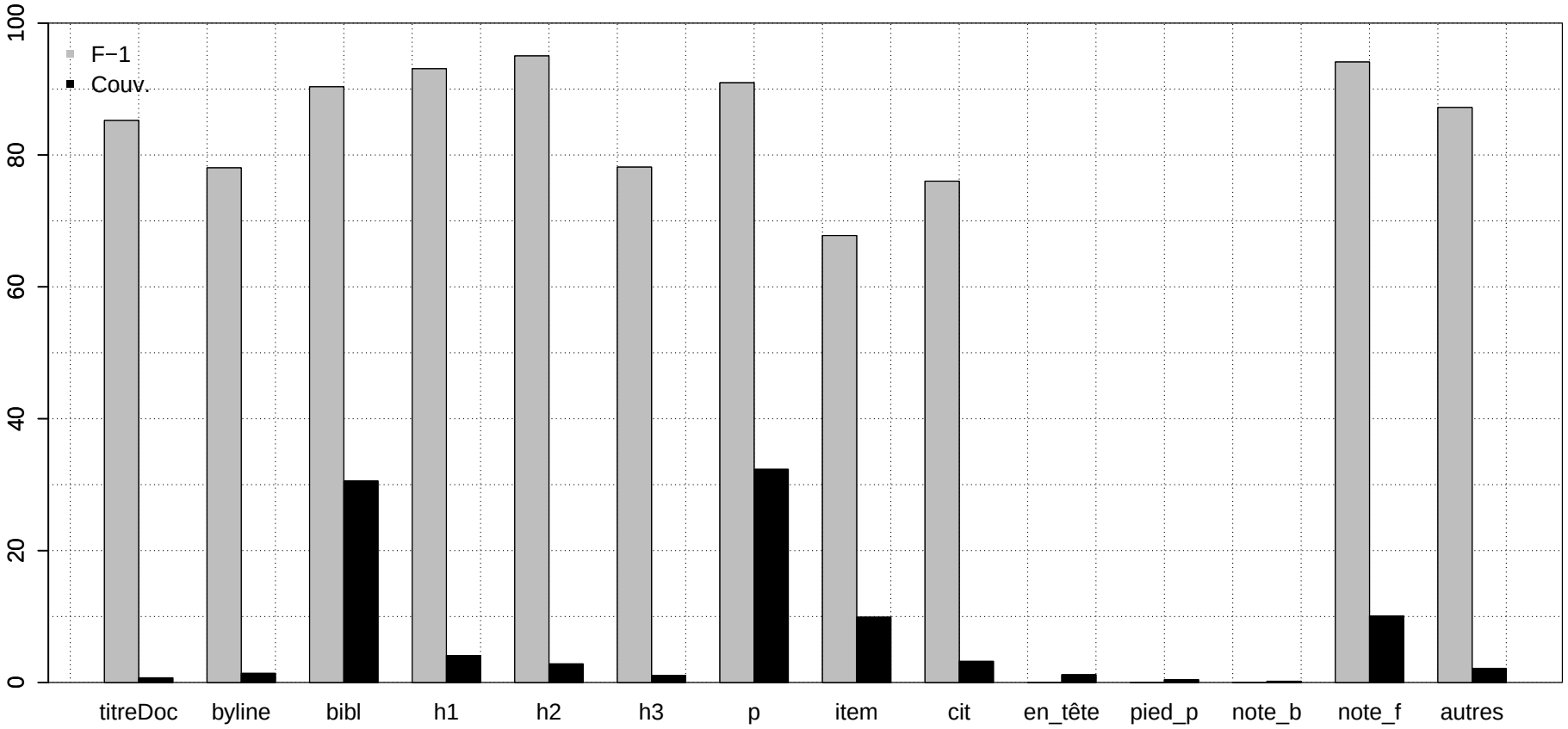
→ Distributions pour LING et GEOP :

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

	LING (25 doc.)		GEOP 21 (doc.)			
étiquettes	Nombre	Moyenne	Nombre	Moyenne	Total	Couv. %
subordination	714	**28.56**	391	18.62	1105	24.02
coordination	2467	**98.68**	1029	49	3496	75.98
Total	3181	127.24	1420	67.62	4601	100

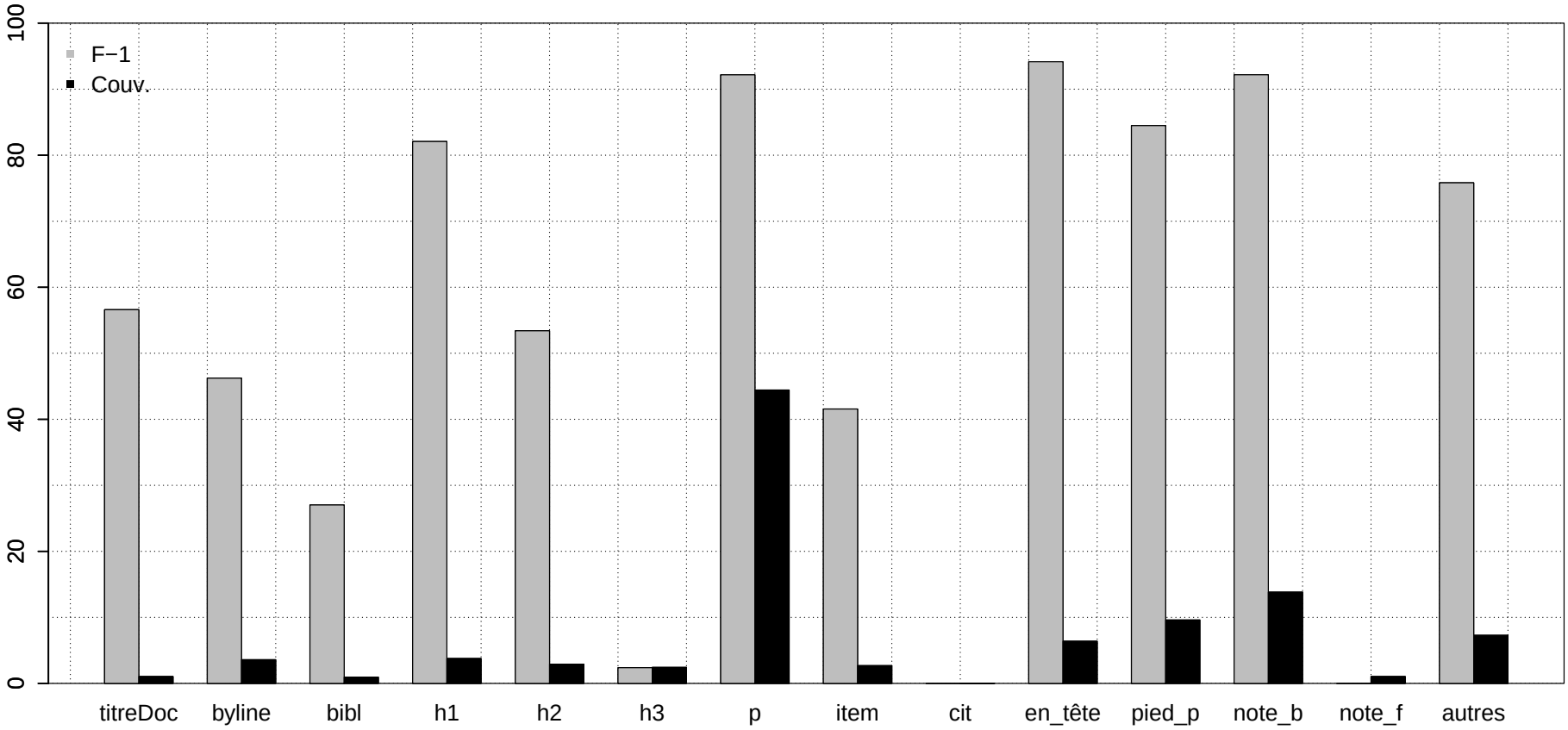
Résultats étiquetage

LING



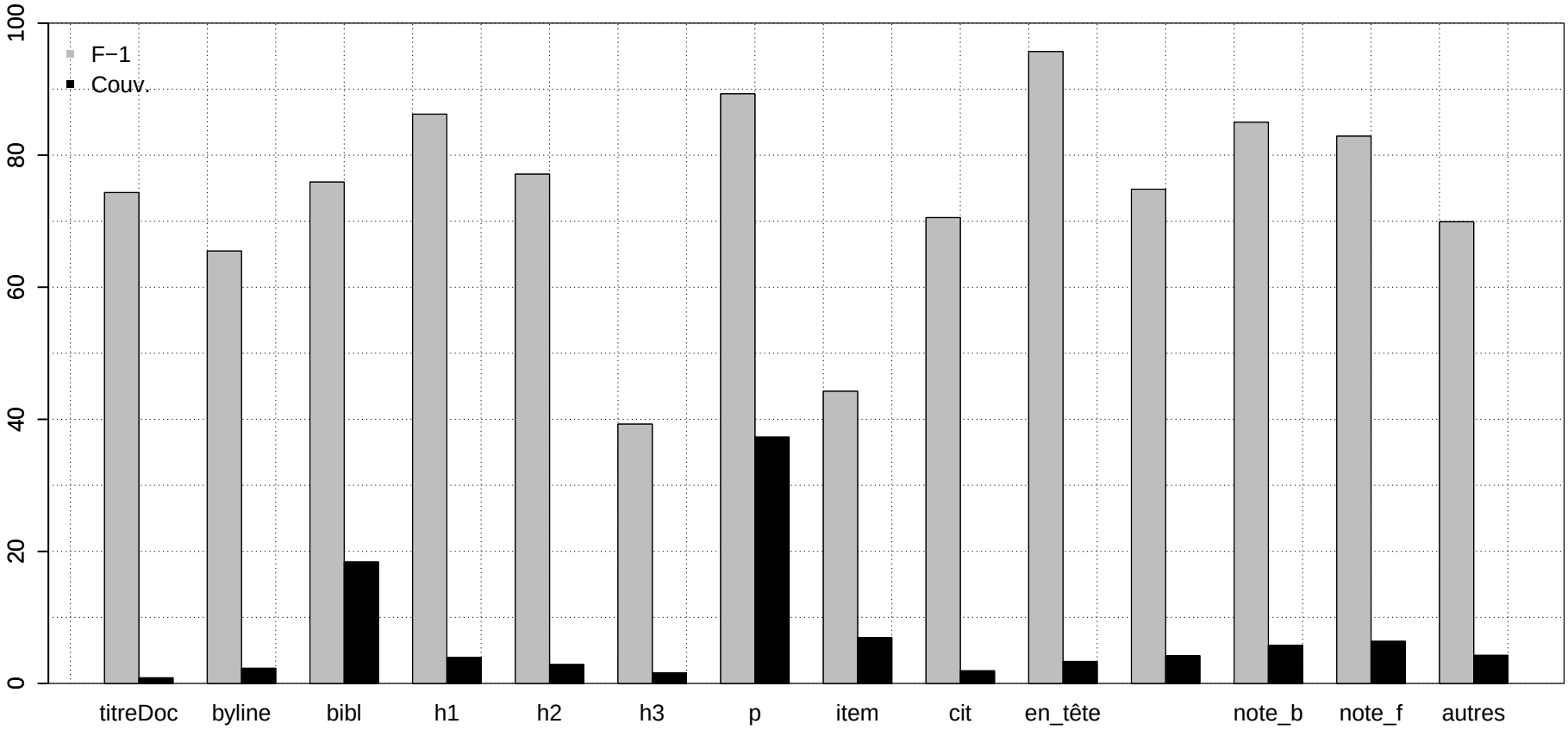
Résultats étiquetage

GEOP



Résultats étiquetage

LING_GEOP



Parsing en dépendances

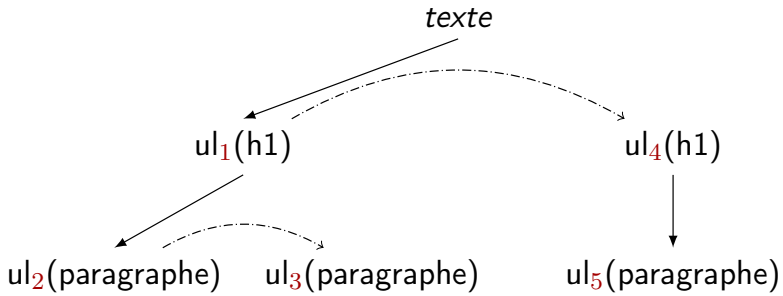
Analyseur LR(1) pour la subordination et la coordination.

1: empiler($texte$, σ)
2: **Tant Que** σ et β non vides :
3: $decision \leftarrow$ choisir_action(σ_0, β_0)
4: **Si** $decision ==$ subordination : /*reduce_subordination*/
5: $D \leftarrow D \cup (\sigma_0, sub, \beta_0)$
6: empiler(défiler(β), σ)
7: **Sinon Si** $decision ==$ coordination : /*reduce_coordination*/
8: $D \leftarrow D \cup (\sigma_0, coord, \beta_0)$
9: dépiler(σ)
10: empiler(défiler(β), σ)
11: **Sinon** /*shift*/
12: dépiler(σ)

Parsing en dépendances

Exemple d'analyse :

Pour un document :
h1,p,p,h1,p.
Nou voulons obtenir ⇒



	action	pile σ	file β	triplet ajouté à D
0		*texte*	h1, p, p, h1, p	
1	reduce_subordination	*texte*, h1	p, p, h1, p	(*texte*, sub, h1)
2	reduce_subordination	*texte*, h1, p	p, h1, p	(h1, sub, p)
3	reduce_coordination	*texte*, h1, p	h1, p	(p, $coord$, p)
4	shift	*texte*, h1	h1, p	
5	reduce_coordination	*texte*, h1	p	(h1, $coord$, h1)
6	reduce_subordination	*texte*, h1, p	∅	(h1, sub, p)
7	shift	*texte*, h1	∅	
8	shift	*texte*	∅	
9	shift	∅	∅	

Grammaire vs. classifieur statistique

Deux stratégies pour comparer grammaire et classifieur

Stratégies	LING	GEOP	LING_GEOP
Exactitude de l'apprentissage sur l'ensemble des dépendances mal classées par la grammaire	14,54% (16/110)	16,66% (4/24)	14,17% (19/134)
Exactitude de l'apprentissage sur l'ensemble des dépendances correctement classées par la grammaire	99,34% (3051/3071)	99,85% (1394/1396)	99,73% (4455/4467)

Exemple

6.1 Un rayonnement limité de la recherche française

La fin des années 1960 voit exploser le domaine de la linguistique, au-delà du structuralisme. Ruwet importe la grammaire générative ; Gross importe les grammaires formelles puis développe une approche originale du traitement des langues naturelles, sur une base d'inspiration harrissienne. Culioli développe sa propre école avec une forte dimension cognitive, Quemada s'intéresse à la linguistique quantitative, *etc*. C'est aussi le temps des grands projets, le plus emblématique étant le lancement du dictionnaire de la langue française et le lancement conjoint de l'Institut Nationale de Langue Française (INaLF) à Nancy.

On assiste donc à un double mouvement.

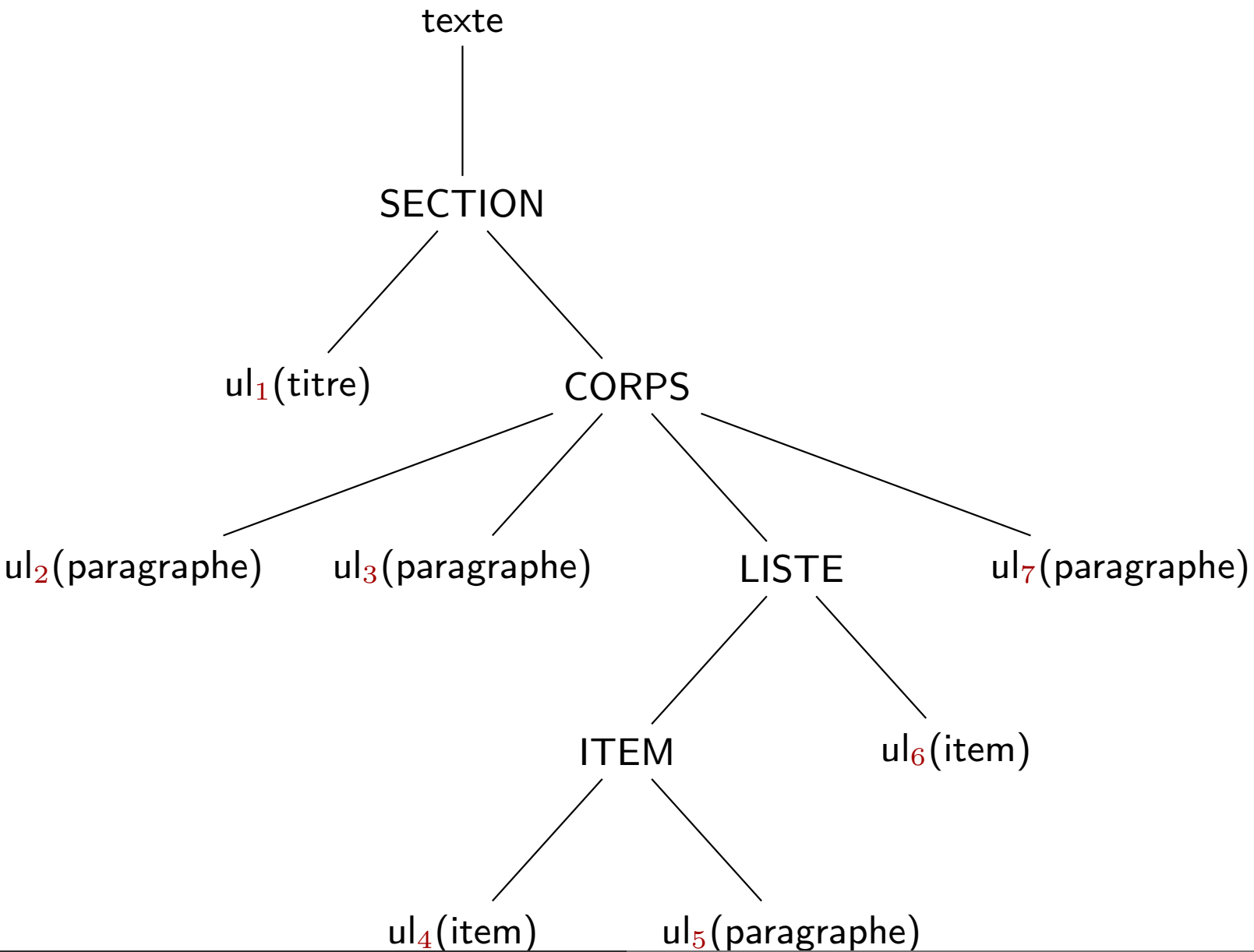
- D'une part l'importation de théories élaborées à l'étranger, le plus souvent aux Etats-Unis, rend moins originales les recherches menées en France. Les chercheurs français traiteront ces théories d'origine anglo-saxonne avec un point de vue original, à l'image du regard critique de Milner sur la grammaire générative (1989) ou de Gross travaillant sur la base d'une analyse de type harrissien (1975). Toutefois, le point de vue français a une influence limitée au-delà des frontières : la France n'est plus le pays moteur en matière d'innovation et de création en linguistique.

  On assiste donc au développement d'écoles françaises sur la base de théories étrangères, mais l'écosystème linguistique français a des interactions limitées avec le monde extérieur. Par exemple, Harris développe à partir des années 1960 sa théorie des sous-langages sur une base distributionnelle mais les recherches de Gross restent relativement hermétiques à ces développements. Les deux chercheurs mènent dès lors des voies séparées et l'influence de Gross restera limitée. Quemada développe de son côté l'analyse lexicographique à partir de comptages systématiques sur corpus, mais ses recherches se développent indépendamment du monde anglo-saxon (même si des représentants de l'école anglo-saxonne ont assisté au grand congrès fédérateur organisé à Besançon en 1961, cf. Léon, 2004).

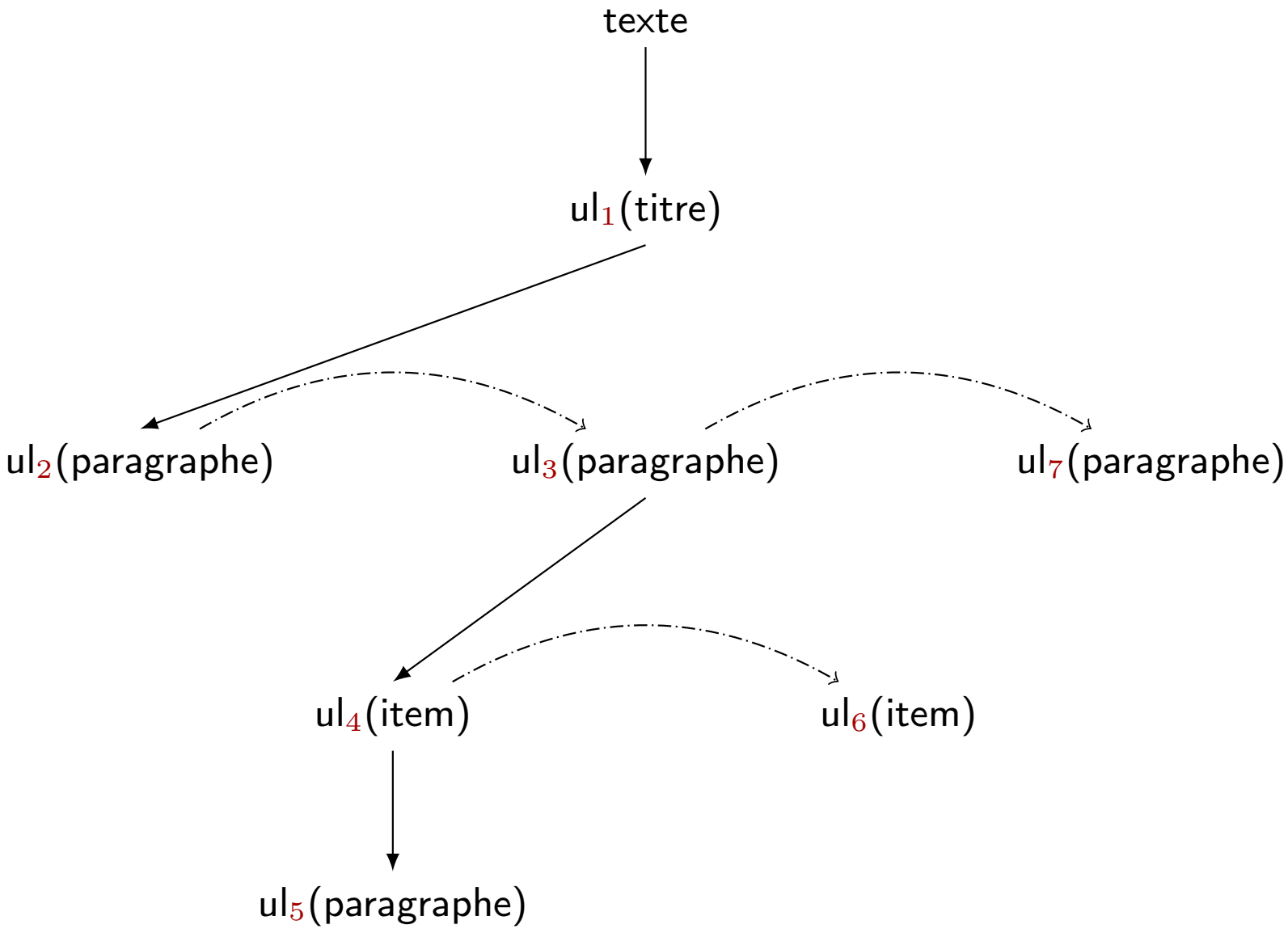
- D'autre part, les projets et les théories propres développés en France ont une audience limitée[6]. Culioli développe sans doute la théorie la plus originale de l'époque mais il écrit peu ; de fait, la théorie culiolienne des opérations énonciatives, qui aurait sans doute pu se développer beaucoup plus largement, reste méconnue à l'étranger et limitée à l'intérieur de la communauté française. Ce n'est que plus tard que les principaux écrits de Culioli seront réellement accessibles et diffusés (Culioli, 1991) mais sans doute trop tard pour être réellement influents sur un plan international.

Le bilan est donc contrasté : alors que la France est en pointe dans les années 1960, les innovations sont principalement le fait d'auteurs anglo-saxons. De fait, la place de la France décline relativement au niveau

Exemple : constituants



Exemple : dépendances



Titres avec une mise en forme d'item

Le système GATT ne répond plus : limites du mercantilisme, évolution des rapports de force, nouveaux acteurs

- *Doit-on « payer » les règles de droit ? La méthode mercantiliste à l'épreuve.*

Depuis l'instauration du GATT, le libre-échange progressait aux rythmes de cycles de négociations, paradoxalement mus par le mercantilisme de l'échange de concessions

(…)

marchés publics et facilitation des échanges – et préférant poursuivre hors de l'OMC, par accords bilatéraux, les deux autres objectifs de régulation. Tous les pays développés avaient, par contre, un point commun : celui de refuser de « payer » par davantage de libéralisation agricole (impliquant des ajustements à coût politique immédiat élevé) l'élaboration de règles de droits (dont le bénéfice économique potentiel se diffuse à moyen ou long terme).

La méthode mercantiliste, issue des négociations du GATT, a rencontré à Cancun ses limites, pour traiter simultanément des enjeux de libéralisation et de régulation.

- *Certains deviendraient-ils aussi égaux que d'autres ? Le « consensus censitaire » à l'épreuve*

Lors de la création de l'OMC, les négociateurs pouvaient se référer à deux modèles de gouvernance. Celui de l'ONU, fondé globalement sur le « suffrage universel » et l'égalité des Etats à l'assemblée générale – sous réserve du Conseil de Sécurité – était aussi celui de l'ancien GATT. Celui des institutions économiques et financières de Bretton Woods était par contre fondé sur le « suffrage censitaire », lié au stock de capital détenu. Issus du GATT, qui était resté essentiellement un « club de riches » aux intérêts économiques comparables, la plupart de ces négociateurs admirait l'efficacité du deuxième système.

Exemple Glaïeul

Gladiolus - Glaïeul

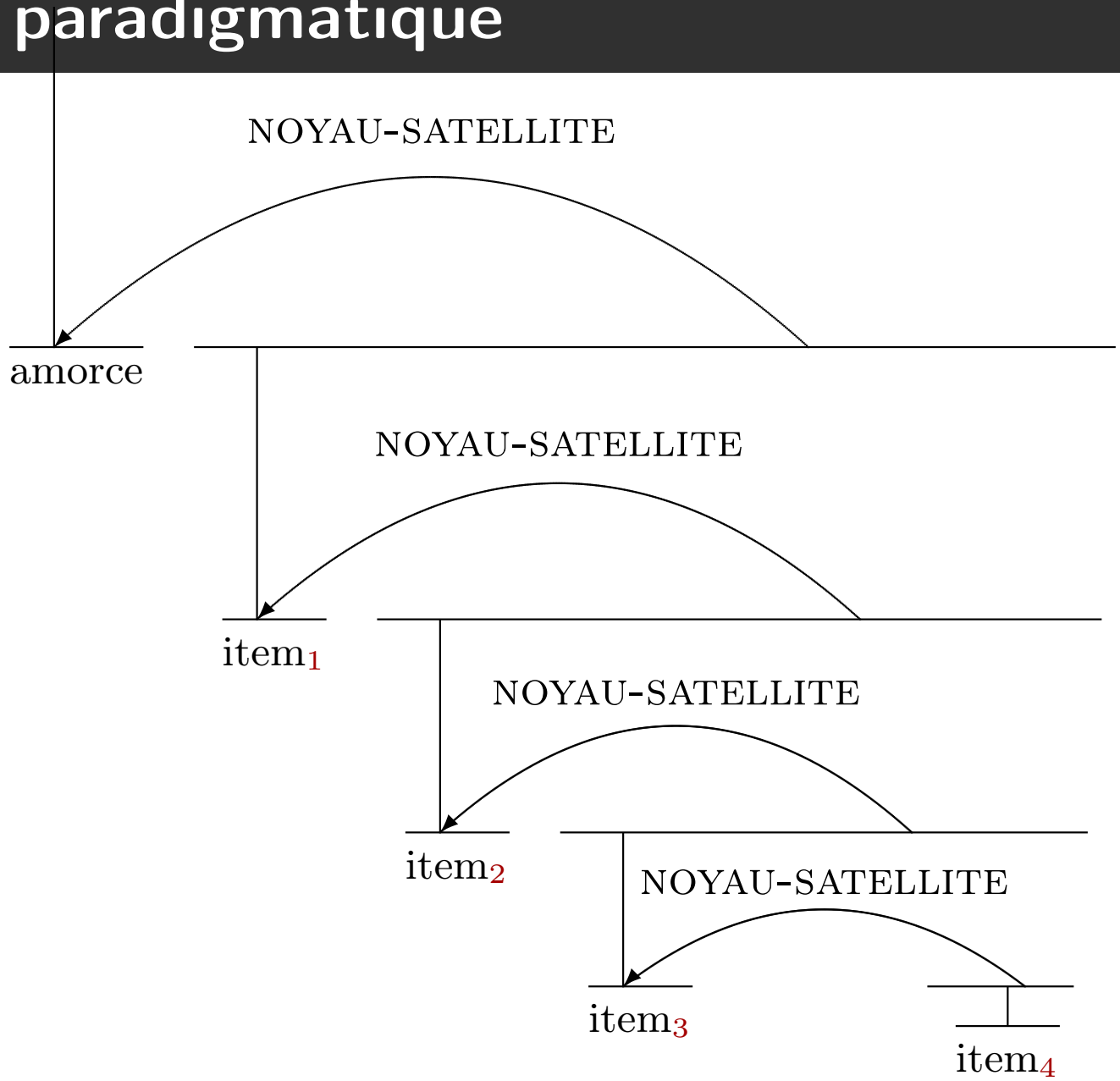
Nom commun : Glaïeul hybride, nommé par les anglophones 'Sword lily'.
Nom latin : *Gladiolus*
Illustration de 'The garden. An illustrated weekly journal of horticulture in all its branches, vol. 30 - 1886 édition William Robinson, contributed by University of Massachusetts Amherst Libraries, U.S.A.
famille : Iridaceae.
catégorie : bulbe à longues racines fragiles et cassantes.
port : élancé, rigide.
feuillage : caduc, vert franc. Deux à trois longues feuilles nervurées engainantes en forme de glaive.
floraison : selon variété au printemps ou de la mi-juillet à la fin septembre durant au moins deux semaines. Epi floral de 6 à 30 fleurons zygomorphes bisexués en forme de trompette à 3 étamines 1 style à 3 aigrettes, visitée entre autre par les abeilles et les colibris. Pour les croisements, féconder tôt le matin un ou deux fleurons par jour à partir du deuxième jour, les fleurons ouverts depuis plus de trois jours ne sont plus féconds.
couleur : dans tous les coloris du blanc au bleu-violet avec une exception le bleu franc et pur, le fond de gorge est fréquemment plus foncé, plus clair ou d'une autre couleur.

fruits : capsules déhicentes à 3 loges contenant chacune une vingtaine de graines ailées.
croissance : rapide.
hauteur : 0.60 à 1.50 m.
plantation : octobre ou novembre pour les glaïeuls de printemps, pour tous à 10-12 cm de profondeur, pointe vers le haut en échelonnant la plantation en 3 fois toutes les 2-3 semaines en les groupant par au moins 10 pour former une belle touffe ou simplement en ligne tous les 15 cm. Prévoir 30 à 40 glaïeuls par 1 m².
Pour les autres, les planter courant mars- avril et jusqu'en juin. Pour la floraison, patienter au moins 10 à 12 semaines (3mois).

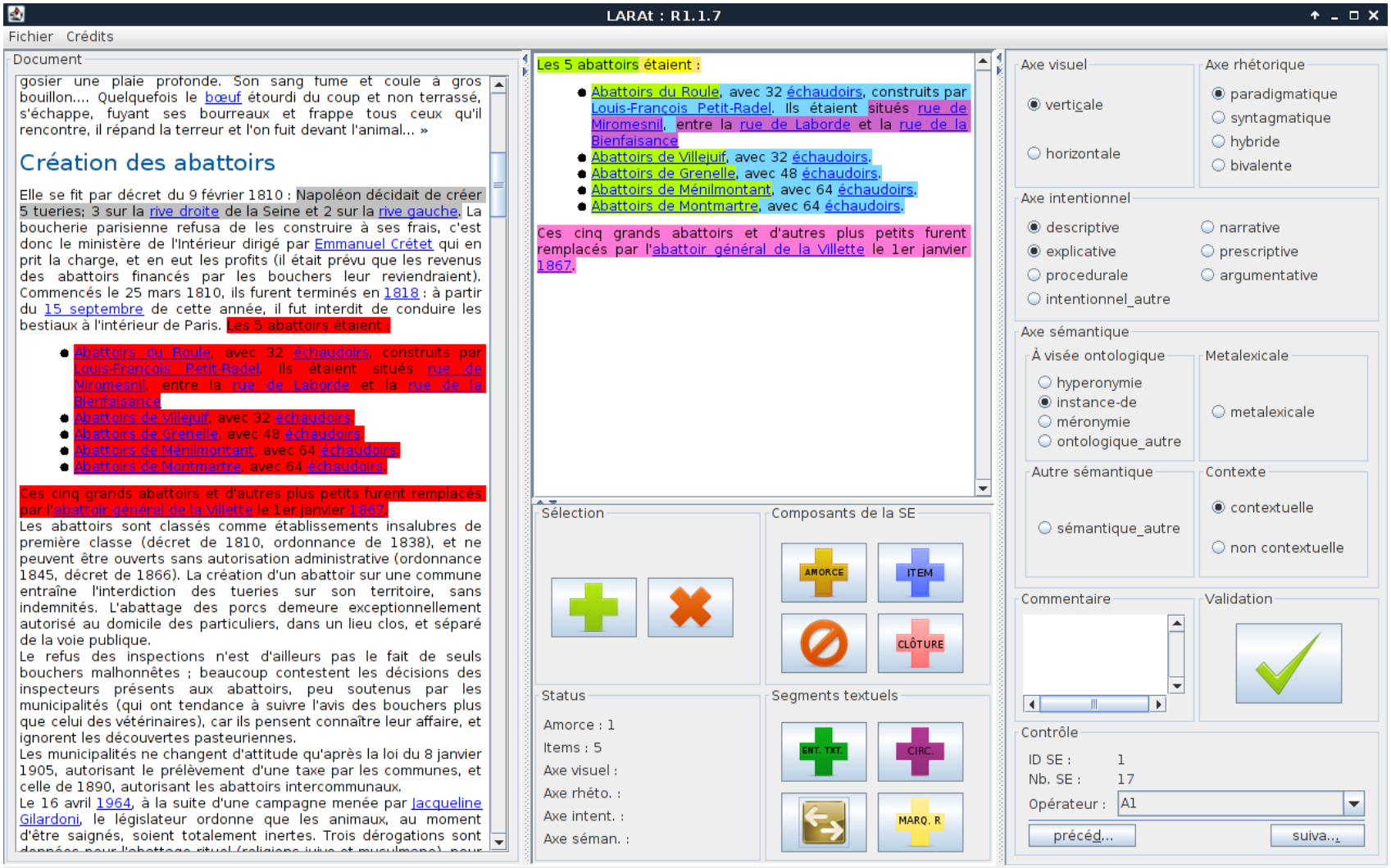
Annexes 2 : Extraction de relations

Annexes pour l'extraction de relations :

SE non paradigmatique



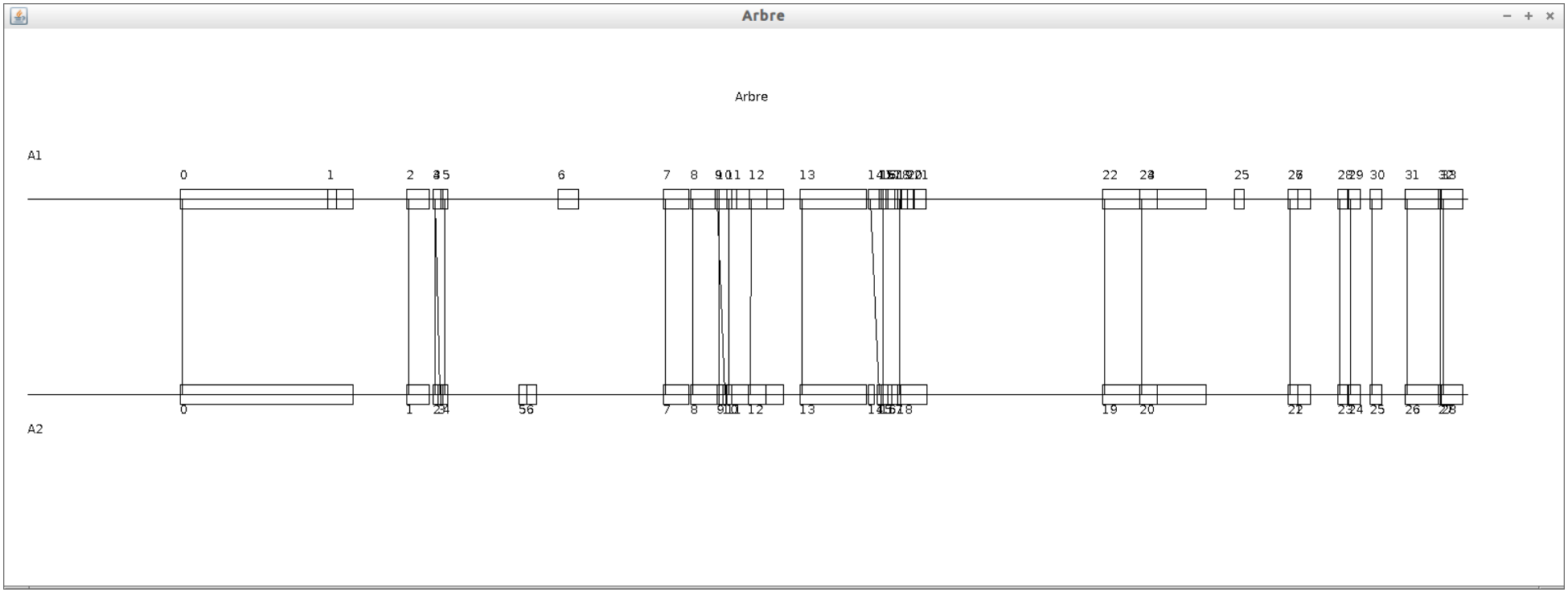
Outil d'annotation LARAt



Outil d'alignement

Nettoyage de l'alignement :

- Exemple pour le document « Arbre ».



Caractérisation sémantique

Caractérisation sémantique (TIA, 2013)

3 types sémantiques de haut niveau :

- **Type à visée Ontologique** : concerne des connaissances du monde
  - **Type hyperonymie**
    ex. voiture et véhicule
  - **Type instance-de**
    ex. Jean Dujardin et acteur
  - **Type holonymie**
    ex. roue et voiture
  - **Type ontologique_autre**
    ex. souris et chat
- **Type metalexicale** : relations qui décrivent le langage
  ex. arête (mathématiques) et arête (ichtyologie)
- **Type sémantique autre** : autres cas.

Voir (TIA, 2013) pour une typologie complète.

Caractérisation sémantique : résultats

Accords et distribution des types sémantiques

Types sémantiques	Kappa κ	Nombre SE	Couverture %
hyperonymie	0,45	268	36,0
instance-de	0,43	196	26,3
holonymie	0,48	39	5,2
ontologique_autre	0,28	42	5,7
metalexical	0,74	149	20,0
sémantique_autre	0,23	51	6,8
Corpus	0,49	745	100

Analyse des traits pour la qualification

Traits	Informations capturées	Composants	corrélation r
t_POS_c	contient : Verbe conjugué	Item	-0,259
t_POS_p	commence par : Déterminant	Item	0,235
t_NbToken	nombre tokens : 5	Item	0,147
t_POS_c	contient : Nom propre	Item	0,132
t_POS_p	commence par : Nom	Item	0,128
t_POS_c	contient : Nom pluriel	Amorce	0,120
t_POS_c	contient : Nom propre	Amorce	0,120
t_POS_p	commence par : Verbe infinitif	Item	-0,113
t_Lexique	marqueurs de relation : *metalexicale*	Amorce	-0,112
t_NbToken	nombre tokens : 3	Item	0,107

Traits pour l'identification des arguments

Traits	Informations capturées.
f_contexte	
t_POS_c	Contexte morpho-syntaxique de l'entité textuelle.
t_Position_c	Position de l'entité textuelle dans l'unité logique.
f_entité	
t_POS_e	Informations morpho-syntaxiques de l'entité textuelle.
t_Inclusion_e	Booléen indiquant la présence d'une inclusion lexicale.
t_NbCar_e	Nombre de caractères de l'entité textuelle.
t_NbToken_e	Nombre de tokens dans l'entité textuelle.
f_document	
t_Logique	Retourne l'étiquette logique de l'unité logique traitée.
t_Position_d	Position d'une unité logique dans l'ensemble du document.
t_Coord_Sub	Informations, pour une unité logique donnée, concernant la présence de coordonnés ou de subordonnés.
t_NbSent_d	Nombre de phrases dans une unité logique.
f_cosinus	
t_Sim_Cos	Similarité cosinus des vecteurs d'entités textuelles.

Analyse des traits pour les arguments

Configurations	Précision	Rappel	$F_{0,5}$-score	F_1-score
Régression logistique	78,98	**69,09**	**76,78**	**73,71**
Similarité cosinus (dim, 500)	**83,71**	30,10	61,72	44,28
Similarité cosinus (dim, 200)	66,52	30,10	53,56	41,45
Baseline	48,37	69,09	51,46	56,91

Algorithme A*

A* pour la recherche du chemin de moindre coût dans un graphe acyclique

```
1:  procedure aStar(racine, O)
2:      enfiler(racine, β)
3:      Tant Que β non vide :
4:          T_i^j ← défiler(trier(β))            /*Tri selon le coût estimé*/
5:          Si estAtteint(O, T_i^j) :
6:              retourner(récupérerChemin(T_i^j))   /*Retourne la solution*/
7:          Sinon
8:              Pour Chaque successeur de T_i^j : /*Continue la recherche*/
9:                  T_{i+1}^k ← copier(successeur)
10:                 P ← récupérerChemin(T_i^j) · <T_i^j, T_{i+1}^k> /*Mise à jour*/
11:                 associerChemin(P, T_{i+1}^k)
12:                 enfiler(T_{i+1}^k, β)
```

Limites des analyseurs syntaxiques

Limites des analyseurs syntaxiques :

« dorsales océaniques » vs. « des dorsales océaniques »

dorsales océaniques : ADJ NC

des dorsales océaniques : DET NC ADJ

« randonnée glaciaire » vs. « une randonnée glaciaire »

randonnée glaciaire : VPP ADJ

une randonnée glaciaire : DET NC ADJ

Exemple de SE

Coordination syntaxique

Les principaux actes cultuels sont :

- le sacrifice, la libation, l'offrande et l'éducation ;
- la prière (invocation, louange, demande, etc.) ;
- le chant et la musique ;
- la lecture de textes sacrés le cas échéant ;
- éventuellement la prédication qui a un rôle important surtout dans les religions abrahamiques et le bouddhisme (mais la prédication peut aussi s'effectuer dans le cadre d'une activité missionnaire qui n'est pas liée à un culte proprement dit) ;
- les pèlerinages, processions.

Exemple de SE

Confusion avec d'autres relations sémantiques

- Association des chemins de fer sud-africains (SARA, Southern African Railway Association), qui représente :
  - CFB (Chemin de fer de Benguela en Angola)
  - Botswana Railway
  - CFM (Chemins de fer du Mozambique)
  - Malawi Railway
  - Central East African Railway in Malawi
  - TransNamib
  - Swaziland Railway
  - Tazara (Tanzania/Zambia Railway Authority)
  - Zambia Railway
  - Tanzania Railways Corporation
  - NRZ (National Railways of Zimbabwe)
  - Beitbridge Bulawayo Railway
  - Metrorail d'Afrique du Sud
  - Spoornet (Afrique du Sud)

Exemple de SE

Noms sous-spécifiés

Noms abstraits qui sont pauvres sémantiquement (Rebeyrolle et Péry-Woodley, 2014)

Le choix de la vigne à planter dépend de plusieurs facteurs :

- la nature du sol ;
- l'exposition ;
- le climat (ensoleillement et précipitations annuelles) ;
- le type de cépage.

Exemple de SE

L'Aquaculture en France

- La France a une tradition ancienne (plus de 1000 ans) de pisciculture extensive en étangs (Limousin, Dombes et nombreux viviers créé par les moines, et utilisation extensive des retenues de moulins dont les vers de farine et déchets de meunerie alimentaient les truites et d'autres poissons ainsi sédentarisés). Au début du XXe, environ 6 000 exploitants d'étangs déclarés, surtout localisés en Région Centre et Rhône-Alpes et Lorraine ont livré 12 000 tonnes (6 790 pour le repeuplement et 2 570 pour la consommation) de carpe, gardon, brochet et tanche, pour un chiffre d'affaires d'environ 16 millions d'euros.
- La salmoniculture en rivière puis la pisciculture marine sont plus récentes. 60 000 tonnes de poissons étaient produites par an au début des années 2000 (en 2002), pour environ 222 millions d'euros de chiffres d'affaires. salmoniculture (133,8 millions de chiffres d'affaires) a permis de produire environ 41 000 tonnes de truites arc-en-ciel (Bretagne et Aquitaine surtout).
- La conchyliculture (huîtres, moules et coquillages) s'est fortement développée sur la façade atlantique.
- les conchyliculteurs ont produit 90 300 tonnes d'huîtres, 4 100 tonnes d'autres coquillages, produites par 52 600 concessions sur le domaine public sur 18 100 hectares et 1570 km de littoral.

Références I

Jacques André, Richard Keith Furuta, et Vincent Quint. *Structured documents*, volume 2. Cambridge University Press, 1989.

Jacques André, Vincent Quint, et al. Structures et modèles de documents. *Le document électronique*, 1990.

N. Asher. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers, Dordrecht, 1993.

Sophie Aubin et Thierry Hamon. Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, pages 380–387. Springer, 2006.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, et Zachary Ives. Dbpedia : A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

John Bateman, Thomas Kamps, Jörg Kleinz, et Klaus Reichenberger. Towards constructive text, diagram, and layout generation for information presentation. *Computational Linguistics*, 27(3) :409–449, 2001.

C. Bush. Des déclencheurs des énumérations d'entités nommées sur le web. *Revue québécoise de linguistique*, 32(2) :47–81, 2003.

Références II

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.

Christiane Fellbaum. *WordNet : An Electronic Lexical Database*. Wiley Online Library, 1998.

Richard Furuta, Jeffrey Scofield, et Alan Shaw. Document formatting systems : survey, concepts, and issues. *ACM Computing Surveys (CSUR)*, 14(3) :417–472, 1982.

Nùria Gala. *Un modèle d'analyseur synthaxique robuste fondé sur la modularité et la lexicalisation de ses grammaires*. PhD thesis, Paris 11, Orsay, 2003.

Marti 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.

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.

William C Mann et Sandra A Thompson. Rhetorical structure theory : Toward a functional theory of text organization. *Text*, 8(3) :243–281, 1988.

Emmanuel Morin. *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. PhD thesis, Université de Nantes, 1999.

Références III

Eugene W Myers. An o(nd) difference algorithm and its variations. *Algorithmica*, 1 (1-4) :251–266, 1986.

E. Pascual. *Représentation de l'architecture textuelle et génération de texte*. PhD thesis, Université Paul Sabatier. Toulouse, France., 1991.

Sylvie Porhiel. Les structures énumératives à deux temps. *Revue romane*, 42(1) : 103–135, 2007.

Richard Power, Donia Scott, et Nadjet Bouayad-Agha. Document structure. *Computational Linguistics*, 29(2) :211–260, 2003.

Cartic Ramakrishnan, Abhishek Patnia, Eduard H Hovy, Gully APC Burns, et al. Layout-aware text extraction from full-text pdf of scientific articles. *Source code for biology and medicine*, 7(1) :7, 2012.

Josette Rebeyrolle et Marie-Paule Péry-Woodley. Énumération et structuration discursive. In *SHS Web of Conferences*, volume 8, pages 3183–3196. EDP Sciences, 2014.

Patrick Séguéla. Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques. *Thèse de doctorat, Université de Toulouse*, 2001.

Assaf Urieli. *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. PhD thesis, Université de Toulouse, 2013.

Références IV

J. Virbel. The contribution of linguistic knowledge to the interpretation of text structures. In J. André, R. Furuta, et V. Quint, editors, *Structured Documents*, pages 161–180. Cambridge Series on Electronic Publishing, 1989.

Dmitry Zelenko, Chinatsu Aone, et Anthony Richardella. Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3 :1083–1106, 2003.