Acquisition de relations sémantiques à partir d'éléments de mise en forme des textes

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

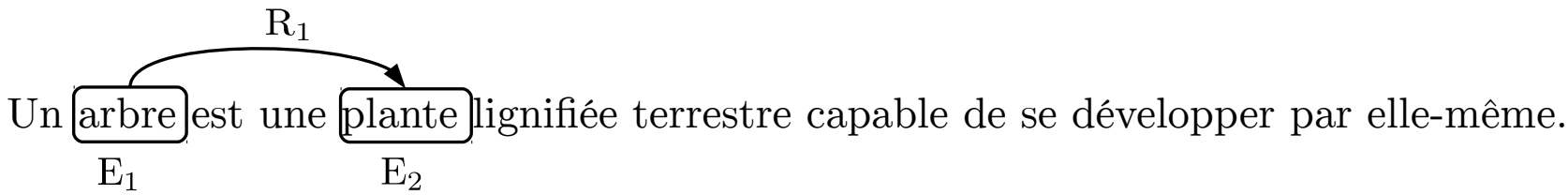
Institut de Recherche en Informatique de Toulouse
Équipe MELODI

21 novembre 2014

Extraction de Relations

- Tâche appartenant à l'extraction d'information (IE) et qui vise à extraire des relations entre des termes.
- 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'hyperonymie 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. : $\textit{arbre} \sqsubseteq \text{plante}$

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

hyperonymie(sérogroupage, test-de-présence-bactérienne)

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

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 termes sont reliés,
  - Identifier la nature de la relation.

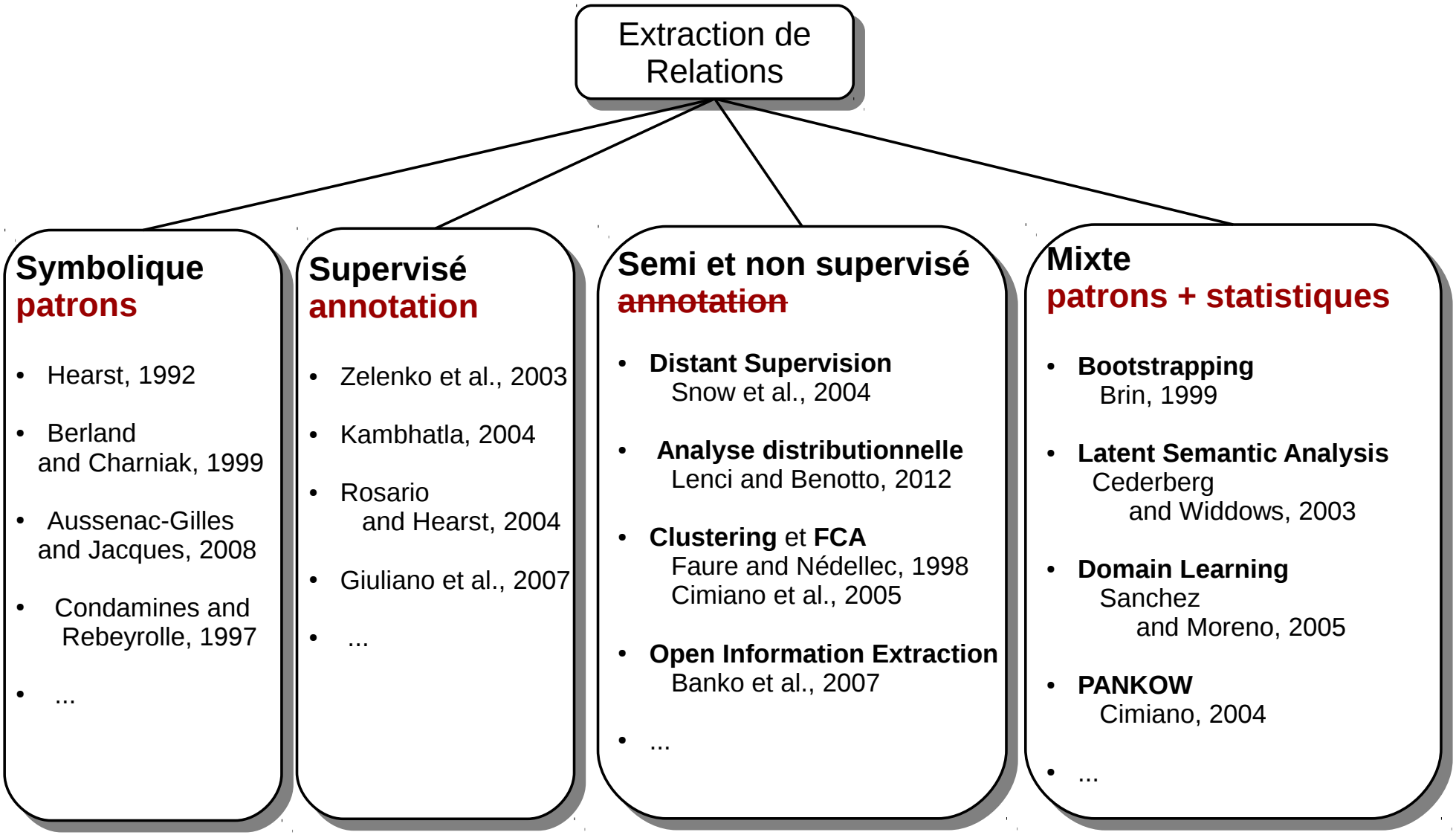
- Traits : contextes des entités, span entre les entités, types des entités, séquence des chunks, arbre de constituants et/ou de dépendances, etc.

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

État de l'art : vue d'ensemble



Problématique :

- Les approches présentées travaillent au niveau de la phrase,
- Deux constats :
  1. Nombreuses relations exprimées au travers de la structure,
  2. Outils actuels pas adaptés à certains segments,
     e.g. : titres, définitions mise en forme, listes à puces, tables, etc.

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. »*

Objectif :

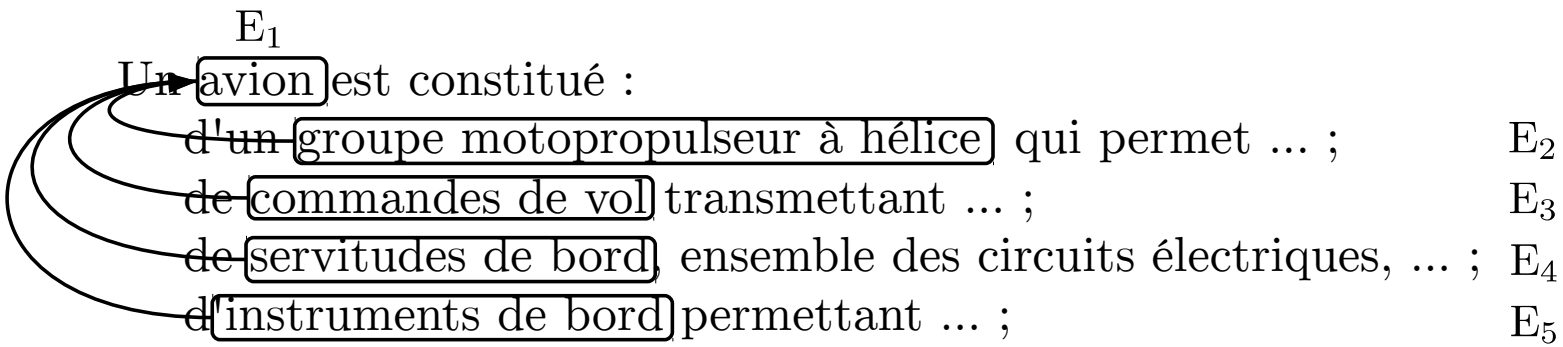
Extraire des relations hiérarchiques au travers de la structure du document.
e.g. : hyperonymie, méronymie.

Hypothèse :

Les éléments énumérés dans une même structure hiérarchique montrent une tendance à partager le même terme classificateur (hyperonymie ou holonymie).

Exemple :

Relation de méronymie entre E_1 et E_2, E_3, E_4, E_5 :



Introduction

Exemples réels : page du CENTAL

> **Projets en collaboration avec des Centres de l'UCL**

- VALIBEL Etiquetage automatisé de corpus transcrits de l'oral. Financement FRFC du FNRS. [2002-2007]
- MOCA
- Collaboration avec le CETIS
- Centre for English Corpus Linguistics (CECL) Collaboration au projet International Corpus of Learner English - ICLE [2002-en cours]
- CEGN Centre d'Études sur Grégoire de Nazianze. [2001-en cours]
- PRLG Projet de recherche en lexicologie grecque. [1990-en cours]

> **Collaboration scientifique interuniversitaire/internationale**

- Projet AdEN: Adolescents et Ecrit Numérique [2008-2013] (Programme ANR, Formes et mutations de la communication : processus, compétences, usages). Le projet AdEN associe cinq équipes de recherche de l'Université de Poitiers, de l'Université de Technologie de Troyes, de l'UCL (Belgique), de la Northern Illinois University (USA) et de l'Université de Valencia (Espagne).
- Projet GALILEI Université Libre de Bruxelles Service de Mécanique Analytique et CFAO
- Projet CONCORLEX : Exploitation linguistique multi-niveaux des ressources textuelles accessibles sur Internet. Université de Marne-la-Vallée.
  Institut Gaspard Monge (IGM) Accord de Coopération CGRI-FNRS-CNRS. [2003-2005]
- Projet IDILL Projet 'Digital Language Learning: An Integrated Perspective': Réalisé dans le cadre du Réseau d'excellence européen Kaleidoscope, ce projet porte sur l'intégration du traitement automatique du langage (TAL) dans le domaine de l'apprentissage des langues assisté par ordinateur (ALAO). Par exemple, une des réalisations de ce projet est le site http://www.idill.org/, site collaboratif présentant des ressources sur le TAL appliqué à l'ALAO. [2006]
- UrgentieAS: Urgentielexicon ArtsenStage Ontwikkeling meertalig medisch lexicon voor artsenstage. Corpusanalyse van geanonymiseerde patiëntenprotocols en medische RSS- corpora. Speciale aandacht voor multidimensionaliteit in lexicale fiches en lexiconstructuur. Koppeling oefenmodules en lexicon op digitale leeromgeving. [2006-2008]

Introduction

Exemples réels : page Wikipédia Volcan

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[5] [réf. insuffisante] ;
- 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 [réf. nécessaire] ; 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 ; Piton de la Fournaise en sont des exemples[6] ;
- 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 éruption le mont Fuji, le Merapi ou le mont Saint Helens[7] ;
- 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 pré comme les Lakagígar ou le Krafla[8].

Comme toute classification de phénomènes naturels, beaucoup de cas sont intermédiaires entre les types purs : l'Etna ressemble à un stratovolcan posé sur un vo 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 types morphologiques. Sans aller aussi loin, on peut ajouter aux précédents deux types très différents[9] :

- 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

Exemples réels : page Wikipédia Arbre

L'arbre procure des matières premières pour un grand nombre d'industries (papetière, seconde transformation c

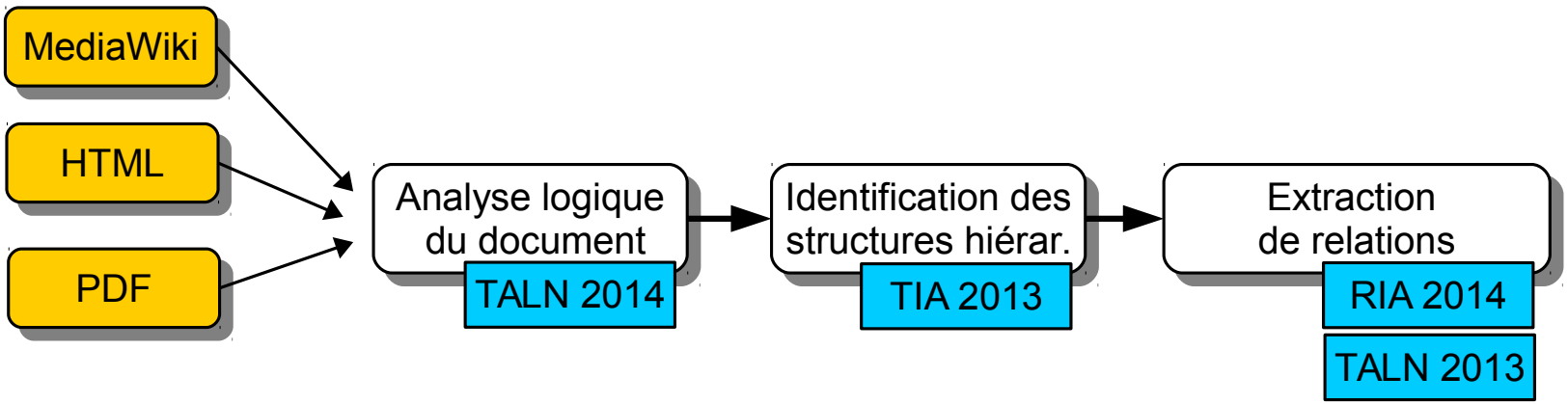
Voici quelques exemples de son exploitation :

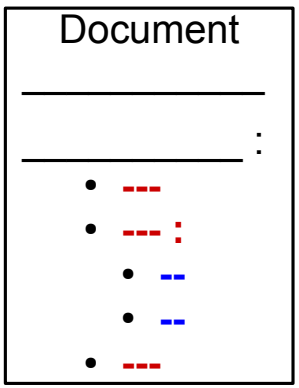
- Pour son bois :
  - bois de chauffage
  - bois d'œuvre (charpente, construction navale…)
  - bois d'ébénisterie
  - cellulose (pâte à papier)
  - charbon de bois
  - tanins (futs de vin en chêne…)
- Pour son écorce (suber) :
  - chêne-liège, chêne rouvre (tan) bouleau quinquina.
- pour ses feuilles :
  - mûrier (élevage du ver à soie)
- Pour ses fruits :
  - fruits frais fruits secs fruits tropicaux.
  - matières oléagineuses : cocotier, olivier, palmier à huile.
  - fibres : kapokier

Deux axes de contribution :

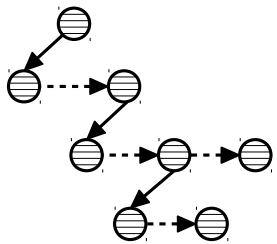
1. Analyse de la structure du document,
2. Extraction de relations dans les structures hiérarchiques.

Trois étapes d'analyse :

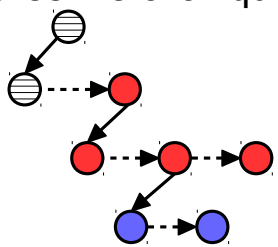




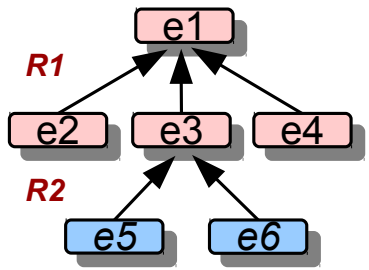
Arbre de dépendance



Identification de structures hiérarchiques



Extraction de relations



1. Introduction

2. Structure du Document
   - Arbre de dépendance
   - Implémentation
   - Expériences sur le PDF

3. Extraction de relations

4. Discussion et Conclusion

Un arbre de dépendance pour la structure du document :

- Ordonne les unités élémentaires,
  e.g. : titres, paragraphes, items, etc.
- Structure intermédiaire entre le niveau logique et discursif.

Propriétés :

1. Construction simplifiée par rapport à une analyse en constituants,
2. Sous conditions, isomorphisme avec la Rhetorical Structure Theory,
3. Facilite l'extraction des structures hiérarchiques.

Contexte :

1. Travail en collaboration avec l'équipe ELIPSE (IRIT),
   Un accès à la mise en forme pour les malvoyants.
2. Nécessité de traiter des documents dans différents formats.

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

1. Titre

P_____
_____
P_____
_____

2. Titre

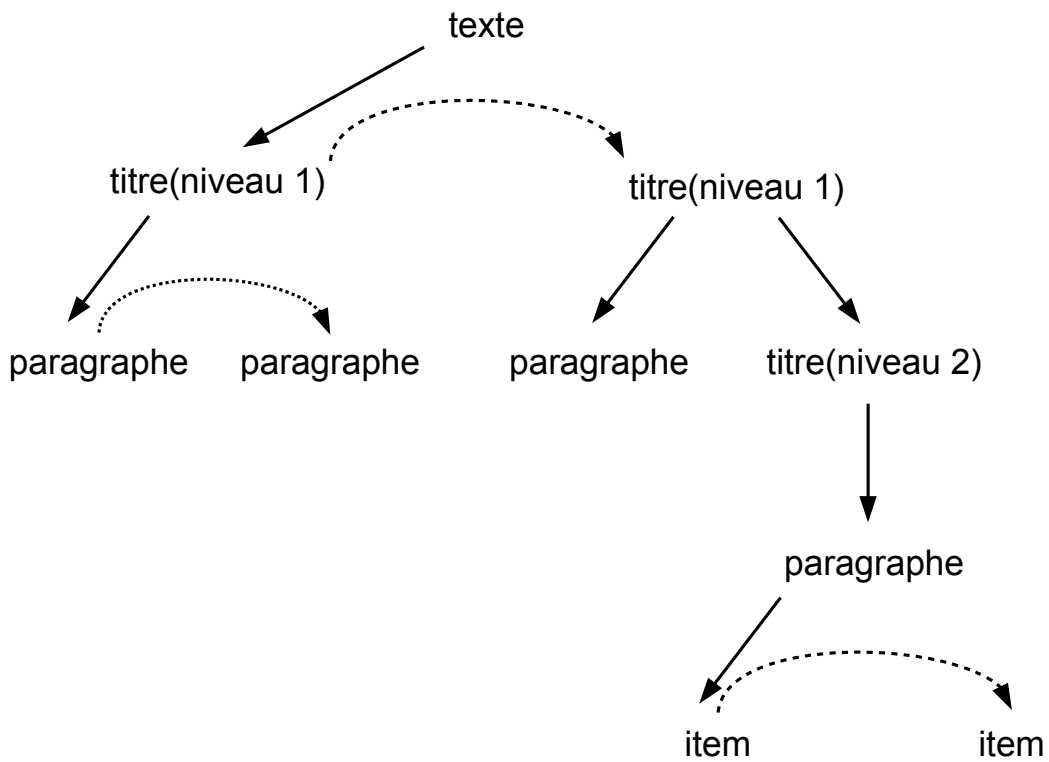
P_____
_____

2.1. Sous-titre

P_____
_____

- item_____
  _____;
- item_____
  _____;

arbre de dépendance



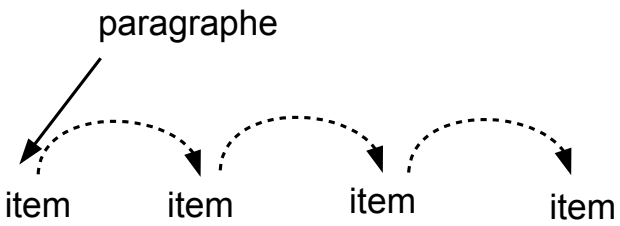
Structure du Document

Intérêt pour l'extraction de relations

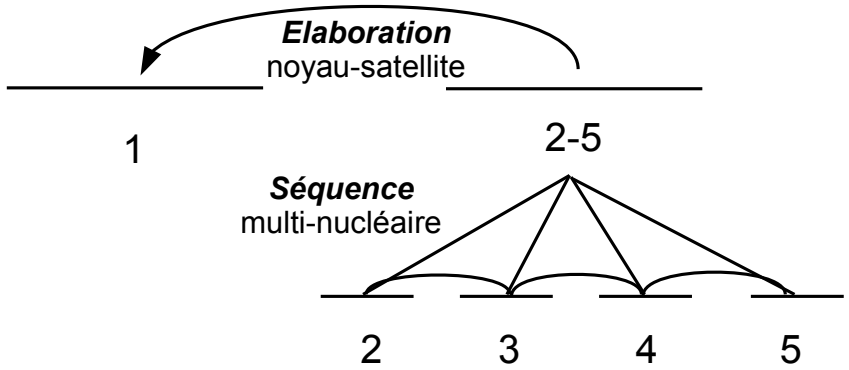
Texte :

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

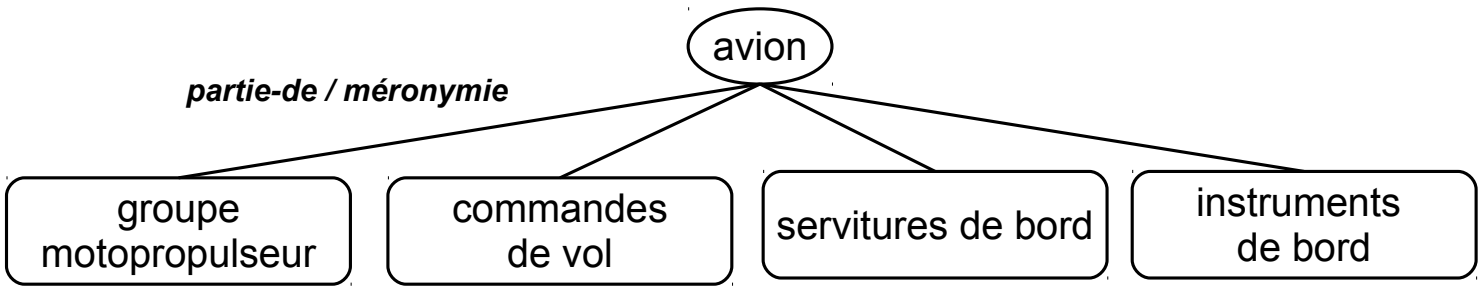
Arbre de dépendance :



Rhetorical Structure Theory :



Représentation conceptuelle :



Implémentation :

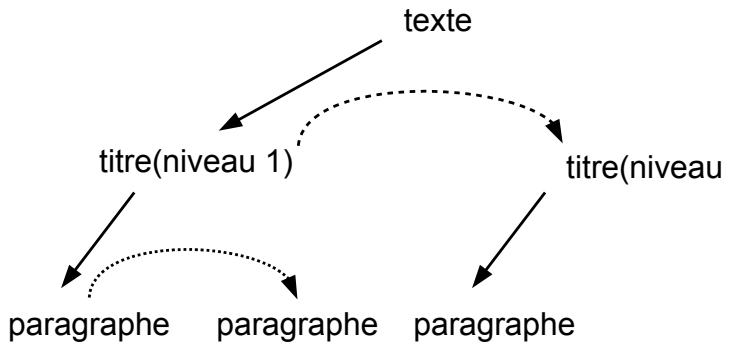
Relier les unités élémentaires par des relation de coordinations et subordination revient à faire un parsing shift-reduce.

algorithm 1 : shift-reduce parsing

```
1:  push root on σ
2:  while β and σ are not empty :
3:      if arc[σ0,β0] == subordination :                    /*reduce*/
4:          a ← β0 and pop β0
5:          push a on σ
6:      else if arc[σ0,β0] == coordination :                /*reduce*/
7:          a ← β0
8:          pop σ0 and β0
9:          push a on σ
10:     else                                                /*shift*/
11:         pop σ0
```

Exemple de parsing :

Pour un document :
h1,p,p,h1,p.
Nous voulons obtenir l'arbre $\Rightarrow$



	transition	pile σ	liste β	dépendances
0		root	h1, p, p, h1, p	
1	*reduce*	root, h1	p, p, h1, p	sub(h1, root)
2	*reduce*	root, h1, p	p, h1, p	sub(p, h1)
3	*reduce*	root, h1, p	h1, p	coord(p, p)
4	*shift*	root, h1	h1, p	
5	*reduce*	root, h1	p	coord(h1, h1)
6	*reduce*	root, h1, p	$\varnothing$	sub(p, h1)
7	*shift*	root, h1	$\varnothing$	
8	*shift*	root	$\varnothing$	
9	*shift*	$\varnothing$	$\varnothing$	

Difficulté variable selon les formats :

- MediaWiki : analyse déterministe,
- HTML : analyse déterministe,
- PDF : analyse non déterministe.

Expériences menées sur le PDF :

Trois étapes d'analyse :

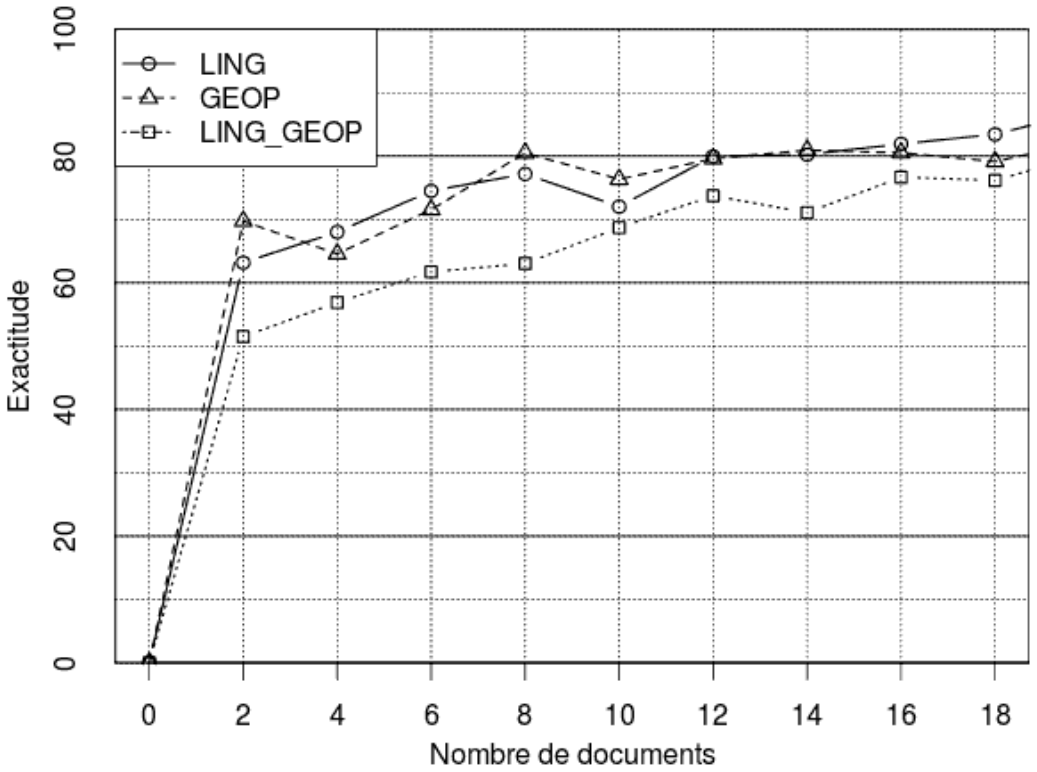
1. Reconnaissance automatique de caractères (OCR),
2. Identification des unités logiques (e.g. : paragraphes, titres, etc.),
3. Construction de l'arbre de dépendance.

Corpus :

1. LING : Articles du CMLF,
2. GEOP : Articles/Rapport de l'IFRI.

1. Identification des unités logiques :

- Utilisation des régularités dans la structure des documents,
- Classifieur : Linear-Chain CRF
- Traits : visuels et séquence.
- Résultats : LING 87,18% , GEOP 82,39%



2. Construction de l'arbre de dépendance :

- Utilisation de l'indentation et de marqueurs de cohésion,
- Classifieur : MaxEnt
- Traits : visuels, lexicaux et parallélisme.
- Résultats : LING 96,41% , GEOP 98,45%

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 chefz de guerre, et sans aulcun delay fit appareillier tout ce qui estoit de besoing pour *aller* en Espaigne *commancer 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,

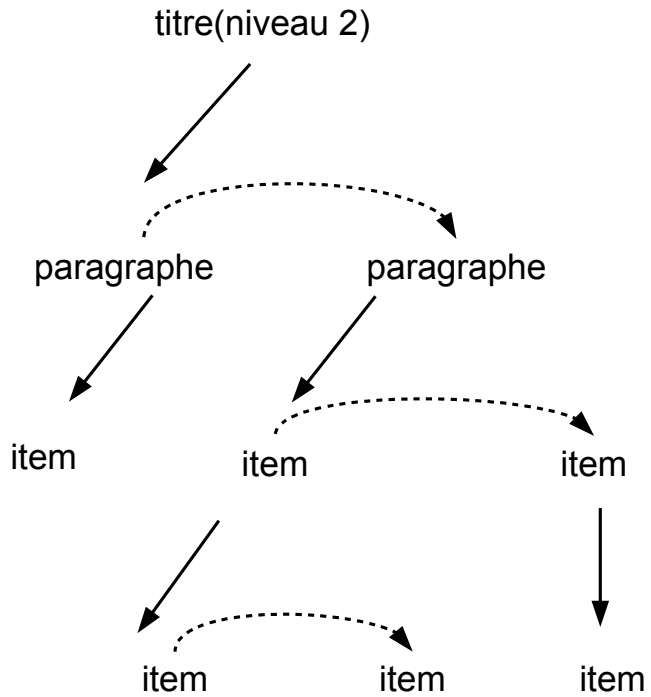
- soit parce que le (co-)texte précise la destination du mouvement:

(4) Nos *alomes* la messe *oïr*; Tuit alomes *vers le mostier*. (*Roman de Renart* 12582, fin 12^{e} – début 13^{e} s.; cité par Littré 1961/62 et Detges 1999: 39)

(5) Il meismes *ala trois serjans apeler* (*Li romans de Berte aus grans pies* XVII, fin 13^{e} 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)



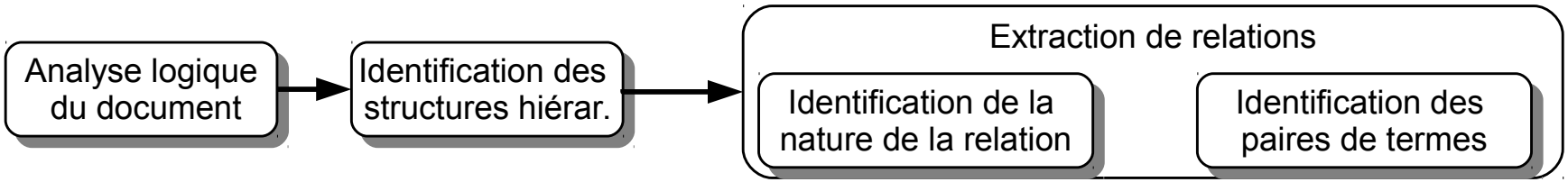
Deux tâches pour l'extraction de relations :

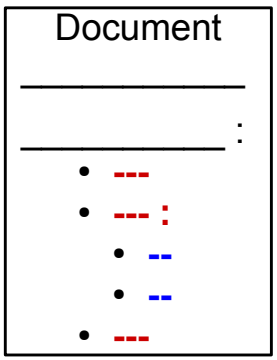
(1) Identifier la nature de la relation,
e.g : *hyperonymie*, *holonymie*, etc.

(2) Identifier les paires de termes,
e.g : « avion », « moteur », etc.

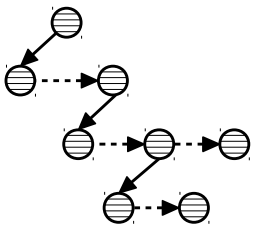
Corpus :

190 pages Wikipédia annotées par trois annotateurs.

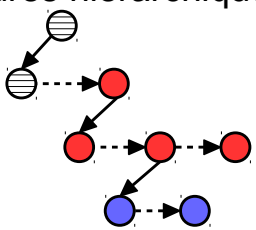




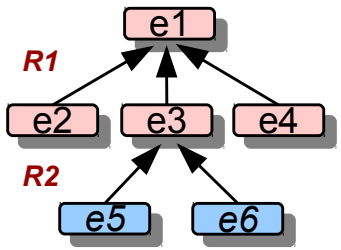
Arbre de dépendance



Identification des structures hiérarchiques



Extraction de relations



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).

1. La compagnie Airbus

Airbus est un constructeur aéronautique international basé à Blagnac...

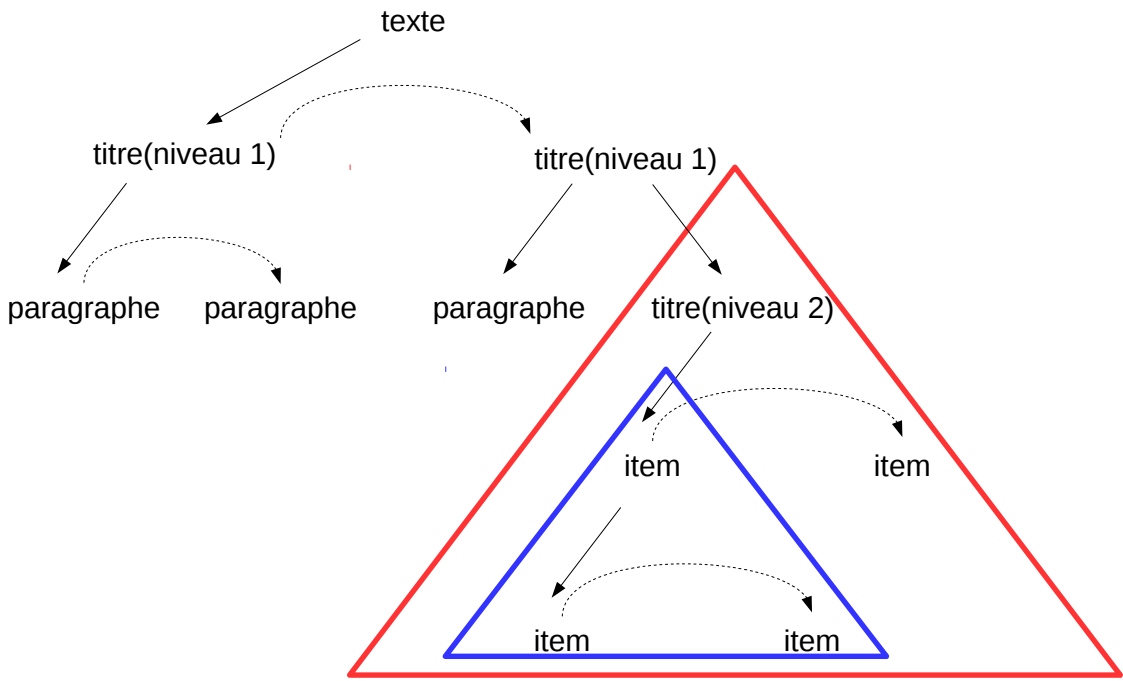
Airbus a produit son premier avion, l'A300, en 1972, et ...

2. L'avion Airbus A380

L'Airbus A380 est un avion de ligne civil très gros-porteur long-courrier quadriréacteur à double pont produit par Airbus civil aircrafts.

2.1. Principaux composants

- Les réacteurs Rolls-Royce Trent 900 :
  - La société Rolls-Royce a été choisie ...
  - Lors du vol inaugural, les réacteurs ont ..
- Les ailes assemblées à Getafe :



Approches précédentes :

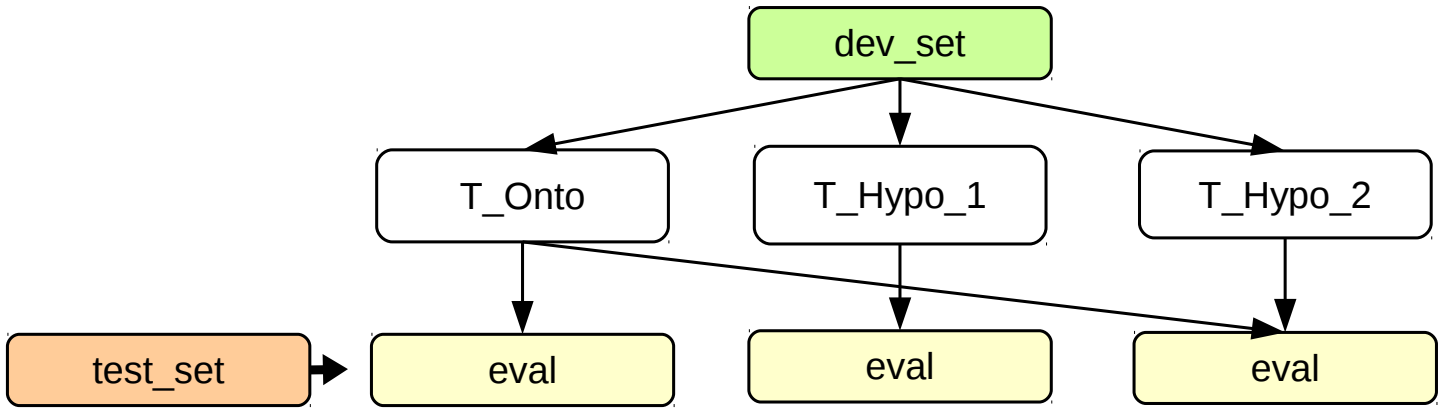
1. Approche symbolique [Kamel and Rothenburger, 2011].
2. Tri-grammes de PosTag et de lemmes :
   Pour chaque tri-grammes de tokens, 2^3 traits :
   Le-DET chat-NC noir-ADJ
   DET chat-NC noir-ADJ
   ...
   DET NC ADJ
3. Classification multi-classes :
   Le même modèle pour tous les types de relation.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

Approche actuelle

- Un modèle binaire par type de relation,
- Utilisation conjointe de modèles,
- Sélection de traits en fonction de la relation.

Trois modèles :

1. **T_Onto** : ontologique vs. non-ontologique.
   hyperonymie, *méronymie*, *autreOntologique* vs. *synonymie*, *autres*, etc.
2. **T_Hypo_1** : hyperonymie vs. non-hyperonymie
3. **T_Hypo_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.

Analyse des traits pour la classe ontologique :

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

	Traits	Segment	*r* 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 d'holonymie	Amorce	0.151
8	Structures hiérar. avec trou syntaxique	Amorce	0.141
9	marqueurs « linguistiques »	Amorce	-0.126
10	nombreTokens=3	Item	0.099

Analyse des traits pour la classe hyperonymie :

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

	Traits	Segment	*r* 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

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	**87.29**	**79.45**
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 ($p < 0.03$).

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	**75.34**
SVM+	74.77	90.22	**81.77**	74.66
baseline	63.01	100.0	77.31	63.01

test_set	ΔF1	ΔExactitude	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.

Tâche 2 :

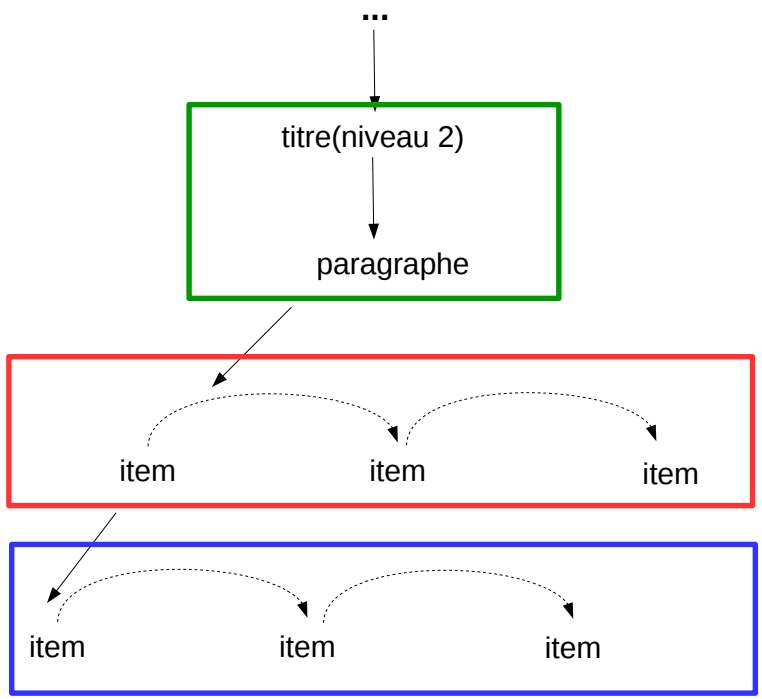
Identification des paires de termes entre l'amorce et l'énumération.

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

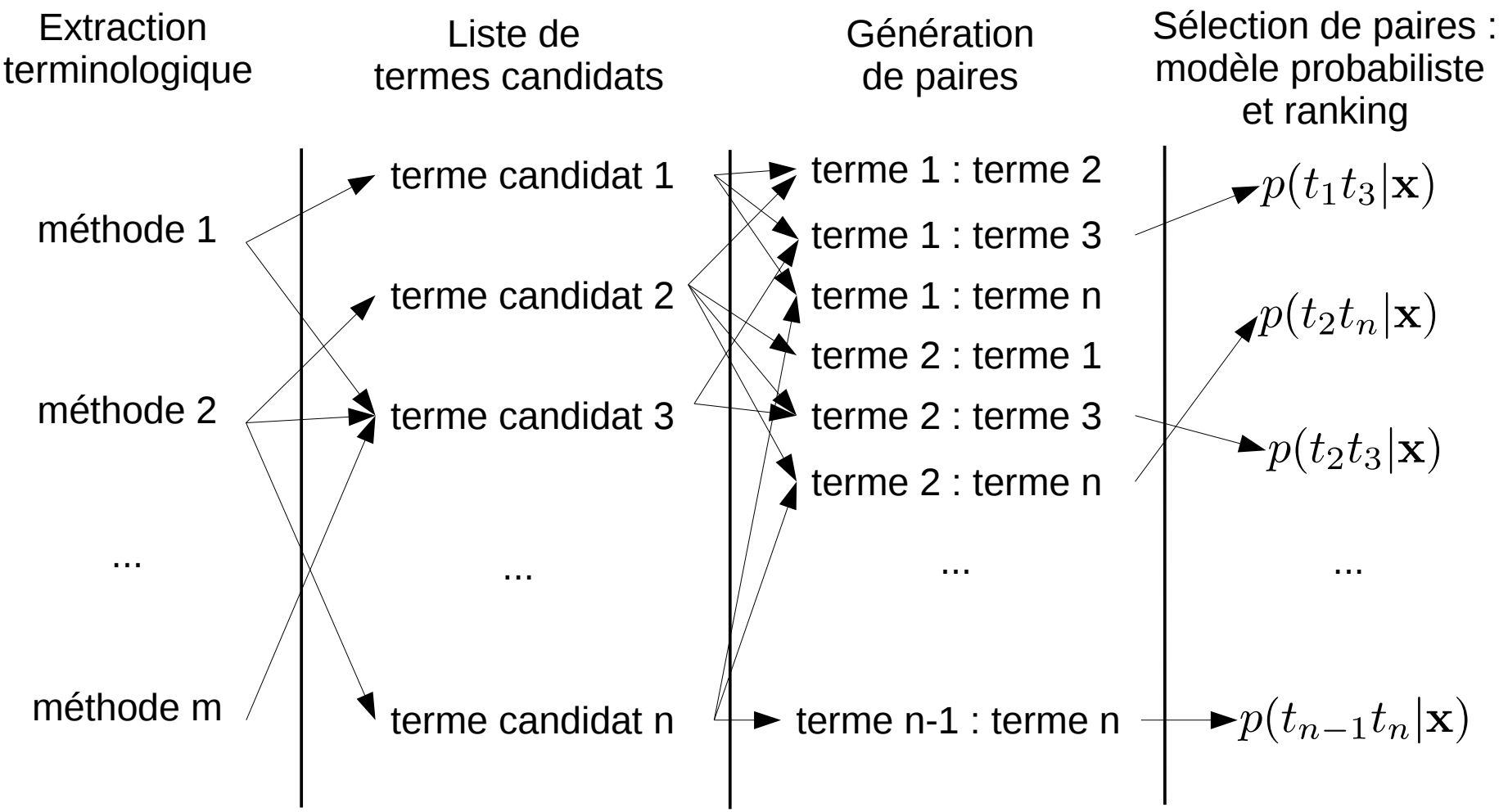
2.6. Atouts liés aux volcans

Par certains aspects, l'homme peut tirer profit de la présence des volcans avec :

- l'exploitation de l'énergie géothermique pour production d'électricité, le chauffage des bâtiments ou des serres pour les cultures ;
- la fourniture de matériaux de construction ou à usage industriel tels que :
  - le basalte qui sert de pierres de construction, de ballast ou de gravas concassé ;
  - la ponce et la pouzzolane qui servent, entre autres, d'isolant dans les bétons ;
  - l'extraction des minerais de soufre, de cuivre, de fer, de platine, de diamants, etc.
- la fertilisation des sols tels les versants de l'Etna qui constituent une région à très forte densité agricole en raison de...

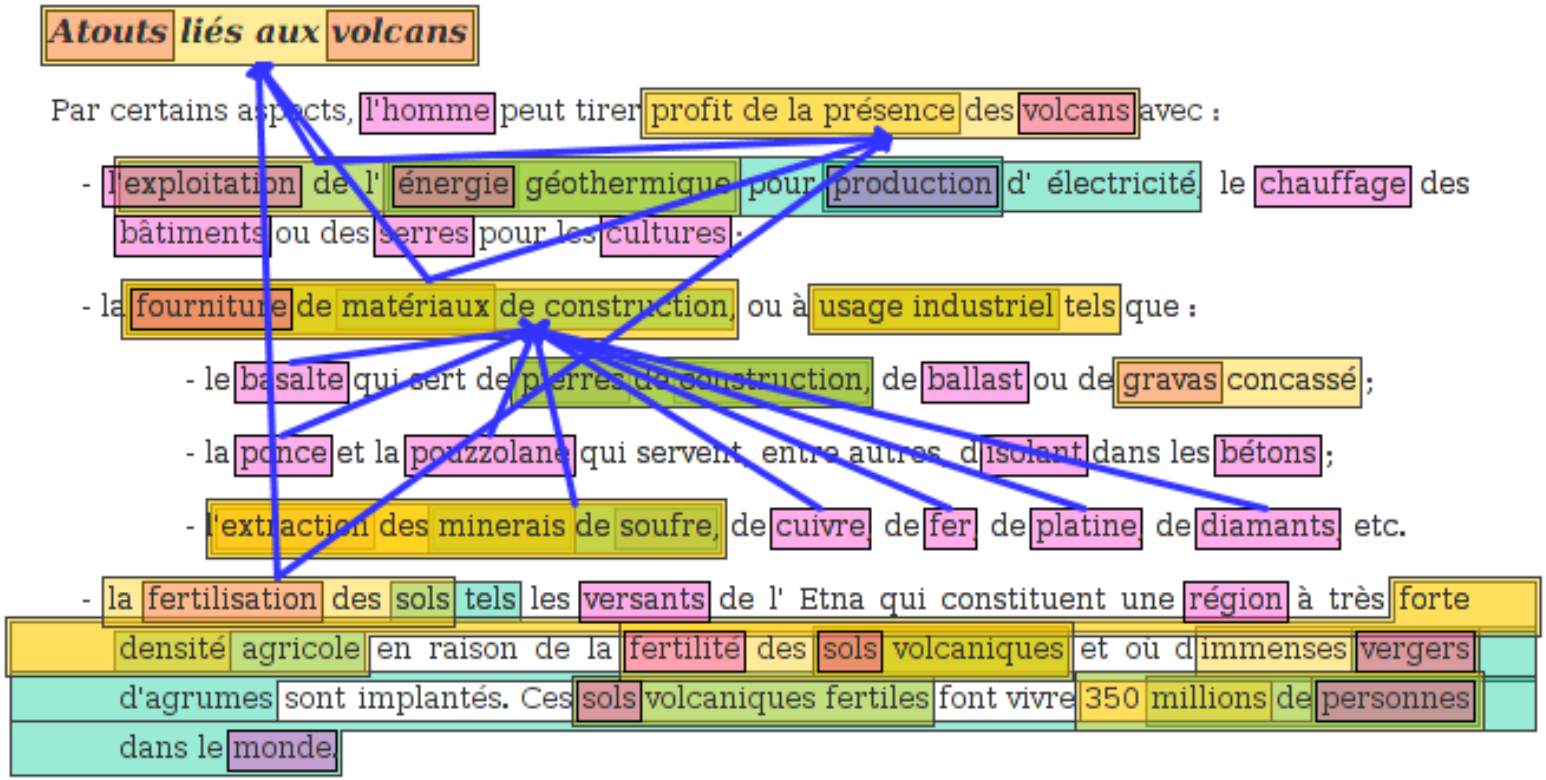


Démarche :


Extraction terminologique
Liste de termes candidats
Génération de paires
Sélection de paires : modèle probabiliste et ranking
méthode 1
méthode 2
...
méthode m
terme candidat 1
terme candidat 2
terme candidat 3
...
terme candidat n
terme 1 : terme 2
terme 1 : terme 3
terme 1 : terme n
terme 2 : terme 1
terme 2 : terme 3
terme 2 : terme n
...
terme n-1 : terme n
$p(t_1 t_3|\mathbf{x})$
$p(t_2 t_n|\mathbf{x})$
$p(t_2 t_3|\mathbf{x})$
...
$p(t_{n-1} t_n|\mathbf{x})$

Exemple :

- Difficulté : espace de recherche assez grand,
  e.g. : pour 80 documents, $\sim$ 100 000 paires,
- Difficulté : plusieurs interprétations possibles



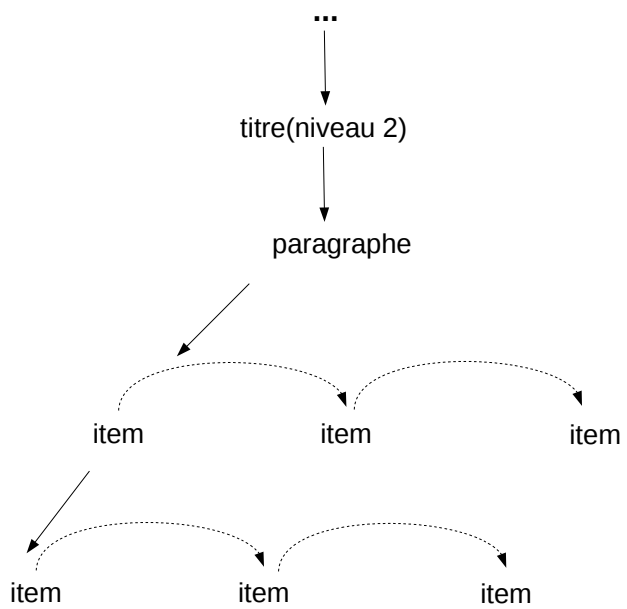
Traits actuels :

- **structure :** changement de niveau, type du chunk, type de dépendance, le subordonnant avec inclusion lexicale, présence de coordonnés, position du chunk dans la structure (premier, dernier), parallélisme de contextes, ...
- **termes :** origine, inclusion lexicale, position, chevauchement, ...
- « **patrons** » : nombres, marqueurs d'énumération (liste de, suivants, etc.), verbes (composer, constituer, etc.).

2.6. Atouts liés aux volcans

Par certains aspects, l'homme peut tirer profit de la présence des volcans avec :

- **l'exploitation de l'énergie géothermique** pour production d'électricité, le chauffage des bâtiments ou des serres pour les cultures ;
- **la fourniture de matériaux de construction**, ou à usage industriel tels que :
  - **le basalte** qui sert de pierres de construction, de ballast ou de gravas concassé ;
  - **la ponce** et la pouzzolane qui servent, entre autres, d'isolant dans les bétons ;
  - l'extraction des minerais de soufre, de cuivre, de fer, de platine, de diamants, etc.
- **la fertilisation des sols** tels les versants de l'Etna qui constituent une région à très forte densité agricole en raison de...



Premiers résultats :

1. Apprentissage : $\sim$ 71 précision pour 235 paires,
2. Baseline : $\sim$ 49 précision pour 2380 paires.

Exemples de paires :

panthéon de rome - temples célèbres

pont-canal de wolfersdorf - ponts canaux

calepitrimerus vitis - parasites animaux

manglier rouge - espèces adaptées

vigne du mont - nom de vigne

marges du glacier - dépôts morainiques*

pratique fédérée - éléments indispensables*

travail - prison **

bloom planctonique - écosystèmes**

Discussion et Conclusion

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

L'inclusion lexicale :

- Trait productif pour l'extraction des paires,
- Perspective : utilisation conjointe avec un parallélisme.

On peut distinguer :
les **digues de protection.** Elles portent alors parfois le nom de levée ...
les **digues de canaux** (d'irrigation, hydroélectriques...), ...
les **jetées**, plus ou moins longues faisant à la fois office de ...
les **ouvrages de protection contre la mer**, de plus en plus ...

Discussion et Conclusion

Apprentissage structuré :

- Apprendre des sets d'hyponymes ou de méronymes.
- Plusieurs idées :
  1. Constitution des sets en sortie et filtrage avec une fonction d'un score,
  2. Graphe de tous les sets possibles et Minimum Spanning Tree.
- D'autres ?

Conclusion :

1. Arbre de dépendance pour la structure du document,
2. Extraction de relations à l'aide de cette structure,
3. Travail en cours avec les paires de termes.

Merci pour votre attention

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 génération

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'inaptitude 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)[15]. 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, québécisme*...), 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.

Approches statistiques : semi et non supervisées

Intuition : intégrer de manière automatique les connaissances linguistiques avec peu ou sans données annotées.

- **Distant supervision** : [Snow et al., 2004], [Mintz et al., 2009]
  1. Projection d'un grand nombre de paires (<-ressources)
  2. Génération d'un grand nombre de traits,
  3. Extraction sur de nouveaux exemples.
- **Analyse distributionnelle** : [Lenci and Benotto, 2012]
  1. Hypothèse d'inclusion des vecteurs de mots,
  2. Extraire les paires selon une mesure de similarité.
- **Clustering** et **FCA** : [Faure and Nédellec, 1998] et [Cimiano et al., 2005]
  1. Extraction des contextes pour tous les triplets $(E_i, R_{i,j}, E_j)$,
  2. Selon similarité des contextes, les regrouper.
- **Open Information Extraction** : [Banko et al., 2007]
  1. Auto-entraînement pour un triplet $(E_i, R_{i,j}, E_j)$,
  2. Extraction de toutes les relations entre entités,
  3. Garder celles qui sont redondantes (ranking).

Approches mixtes : symboliques et statistiques

Intuition : utilisation conjointe de patrons et de mesures statistiques calculées sur le Web.

- **Bootstrapping** : [Brin, 1999], [Agichtein and Gravano, 2000]
  1. Extraction des paires avec des patrons,
  2. Génération de patrons en projetant les paires extraites,
  3. Itération jusqu'à un certain seuil de bruit.
- **Latent Semantic Analysis** : [Cederberg and Widdows, 2003]
  1. Extraire des paires d'entités avec des patrons,
  2. Les filtrer selon la similarité sémantique entre entités.
- **Domain Learning** : [Sanchez and Moreno, 2005]
  1. Recherche Web selon un ou plusieurs mots-clefs,
  2. Extraction par patrons dans les documents retournés.
- **PANKOW** : [Cimiano et al., 2004]
  1. Extraction d'entités et génération de requêtes par patrons,
  2. Recherche Web et filtrage selon le nombre de résultats.

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 noeuds.
  $\rightarrow$ 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 chefz de guerre, et sans aulcun delay fit appareillier tout ce qui estoit de besoing pour *aller* en Espaigne *commancer 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 *oïr*; Tuit alomes *vers le mostier*. (*Roman de Renart* 12582, fin 12^{e} – début 13^{e} s.; cité par Littré 1961/62 et Detges 1999: 39)

(5) Il meismes *ala trois serjans apeler* (*Li romans de Berte aus grans pies* XVII, fin 13^{e} 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»):

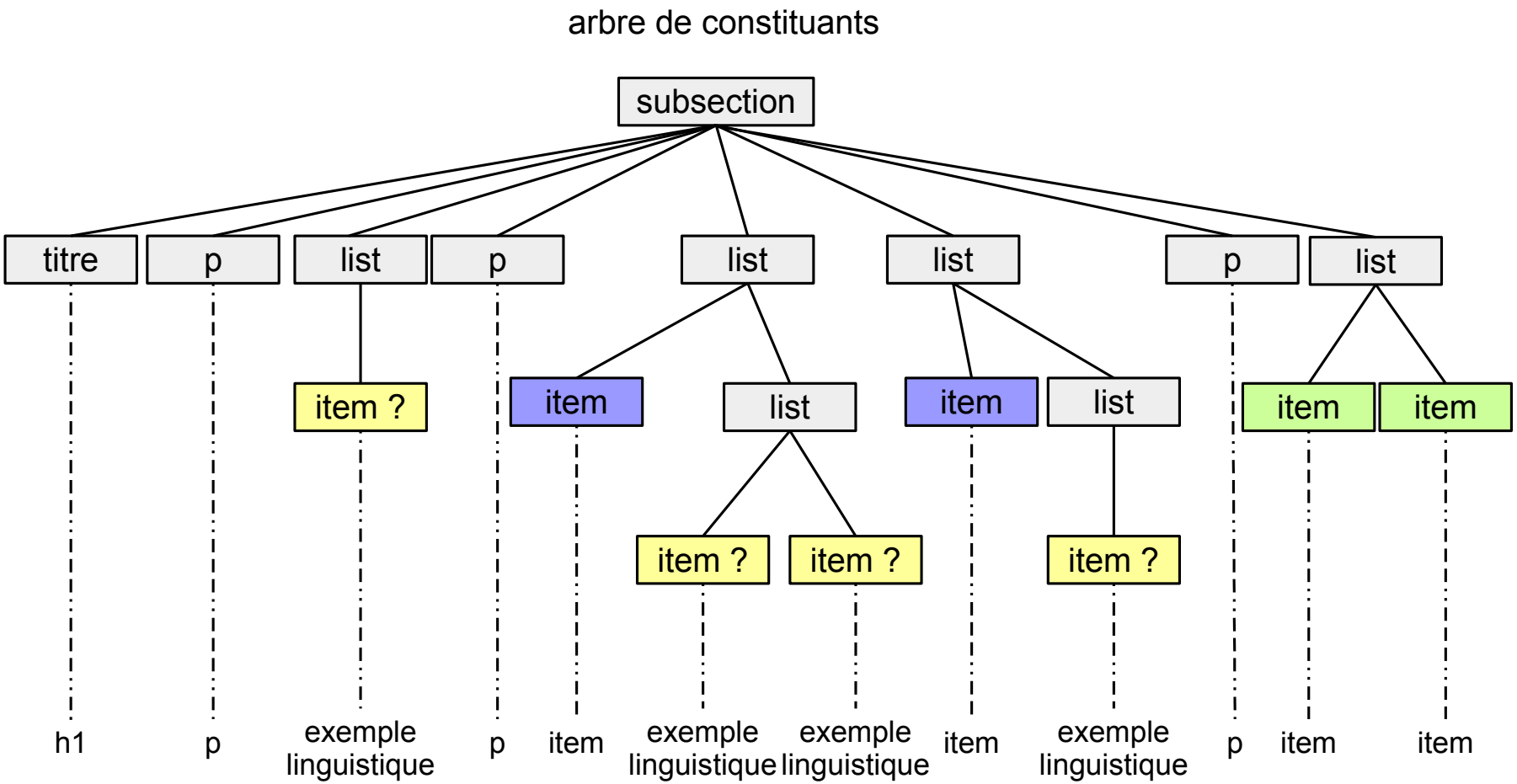
(6) Il est bien temps de deviser / Les personnaiges et nommer. / Je vous les *veulx* nommer à tous. / Je *voys* au Monde commencer. (*Moralité de Charité*, 1532-1550, passage cité par Gougenheim 1929/1971: 98 et Detges 1999: 39)

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,

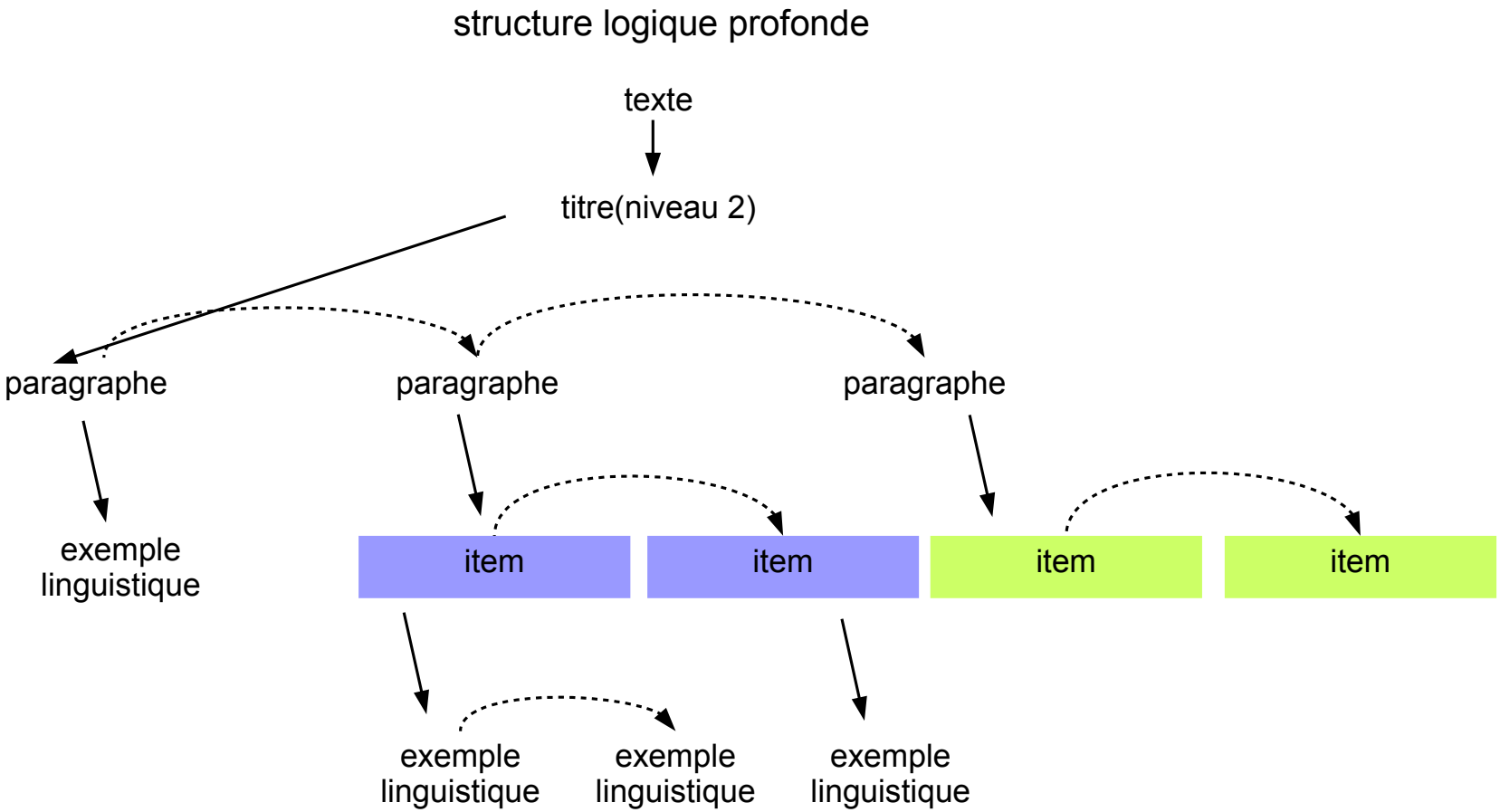
Diapositives supplémentaires

Structure du Document : constituants et dépendances

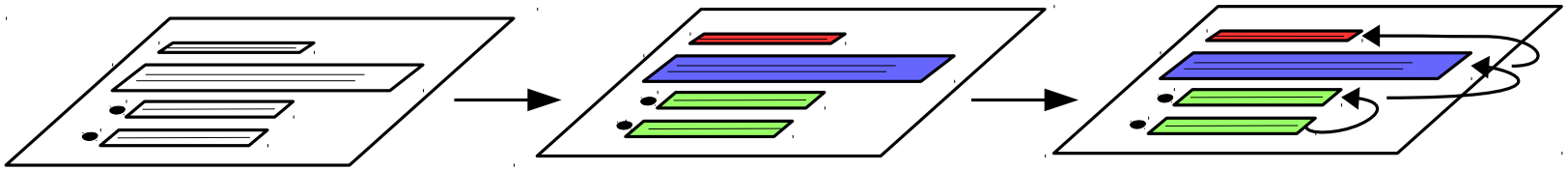


Diapositives supplémentaires

Structure du Document : constituants et dépendances



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>
```

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	pied p.	note p.	bibl	autres
LING	304	1241	380	123	45	16	394	1173	82
Moy.	12,1	49,6	**15,2**	**4,9**	1,8	0,6	15,7	**46,9**	3,2
GEOP	241	1189	72	1	171	257	122	398	195
Moy.	11,4	56,6	3,4	0,05	**8,1**	**12,2**	5,8	18,9	**9,2**

Annotation de l'arbre de dépendance :

1. Ajout de subordination et coordination entre unités logiques
   Grammaire de dépendance
2. Ajout des structures multi-échelles de ANNODIS ME

→ Distributions pour LING et GEOP :

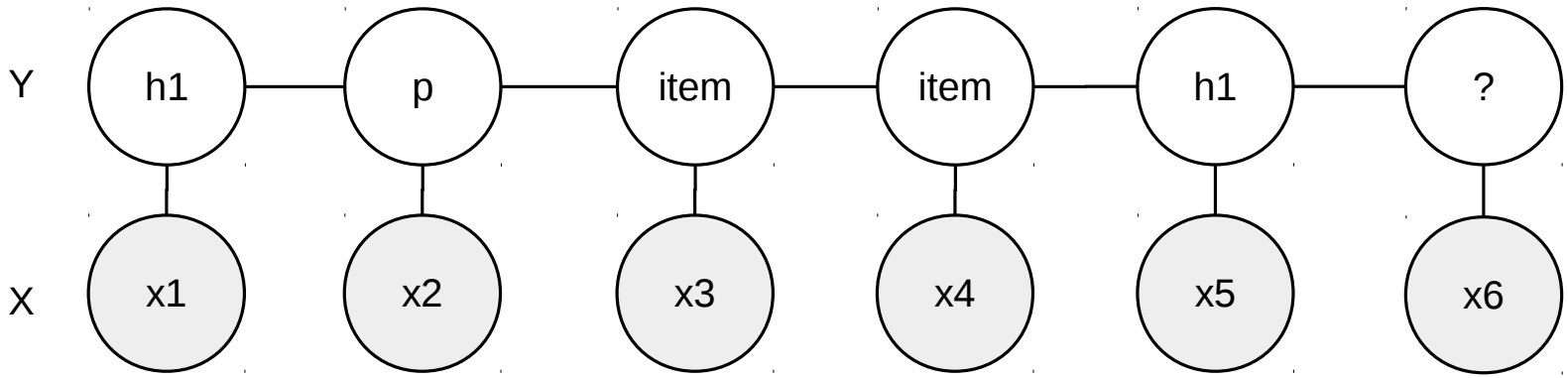
- Nombreuses subordinations et coordinations dans LING
- Prédominance générale des coordinations sur les subordinations

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

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



Traits pour la Tâche 1 :

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

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

Résultats pour la Tâche 1 (exactitude) :

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

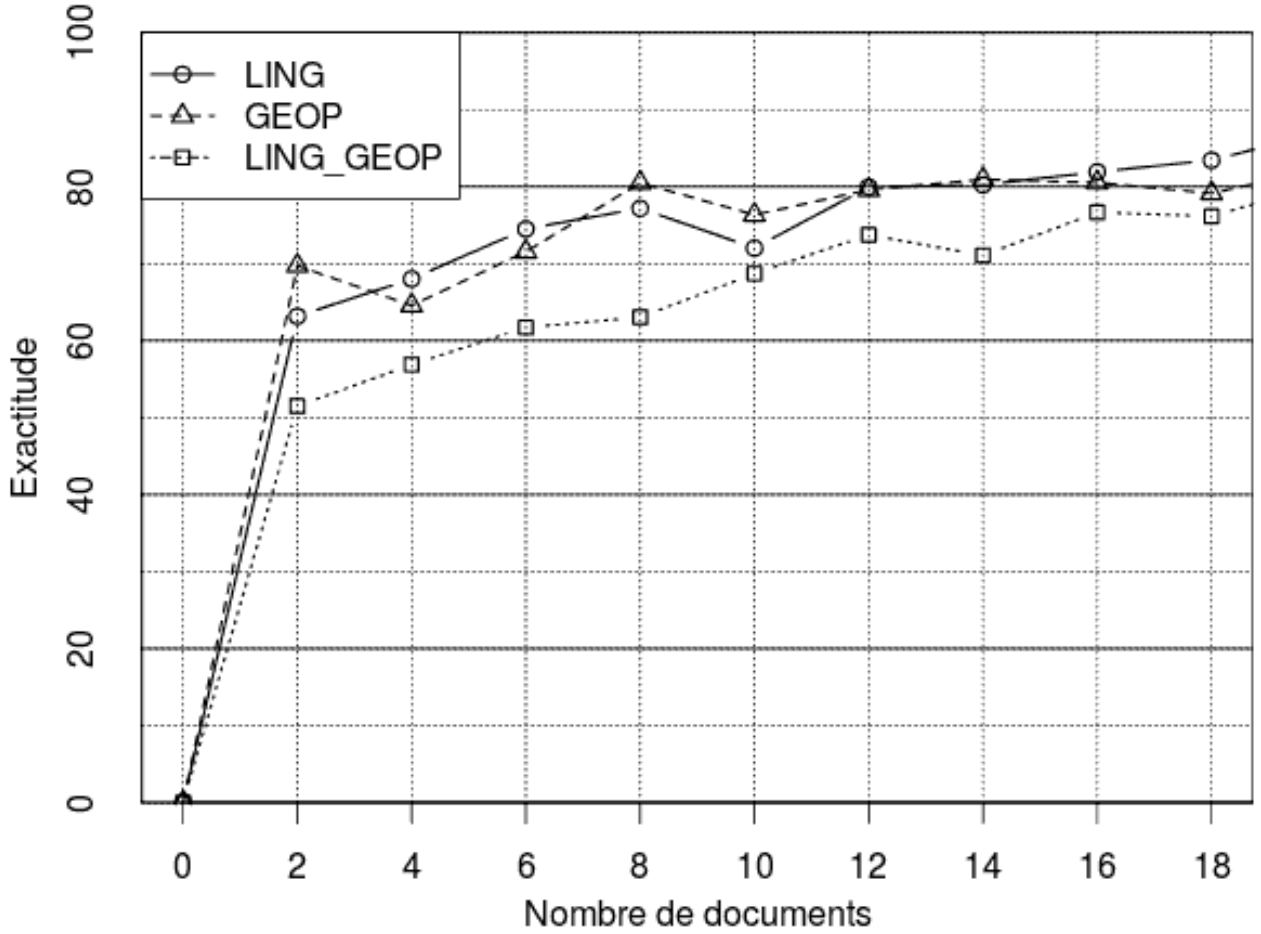
- Baseline :
  - classification en paragraphes (classe majoritaire)
- Traits locaux :
  - LING présente des objets ambigus visuellement (exemples linguistiques)
- Traits de séquence :
  - La séquence de labels permet de gérer la variation locale des unités
  - GEOP présente une structure simple et une 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

Courbes d'apprentissage :

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



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 σ
2: **while** β and σ are not empty :
3: **if** arc[σ_0,β_0] == subordination : /*reduce*/
4: $a \leftarrow \beta_0$ and pop β_0
5: push a on σ
6: **else if** arc[σ_0,β_0] == coordination : /*reduce*/
7: $a \leftarrow \beta_0$
8: pop σ_0 and β_0
9: push a on σ
10: **else** /*shift*/
11: pop σ_0

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
visuels	Présence d'indentation, de tirets, de puces, de « : », etc.
lexicaux	Présence de marqueurs d'intégration linéaire (e.g : *D'une part*, *D'autre part*, etc.)
labels	Paires de labels (e.g : titre-paragraphe, item-item, paragraphe-item, etc.) et égalité de labels
parallélismes	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%	**98,45%**	**97,23%**
Grammaire	**96,54%**	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.

Corpus LARA :

- Débuté en juin 2013,
- Trois besoins :
  1. Des données pour l'apprentissage et l'évaluation,
  2. Une meilleur caractérisation des Structures Hiérarchiques,
  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.

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.

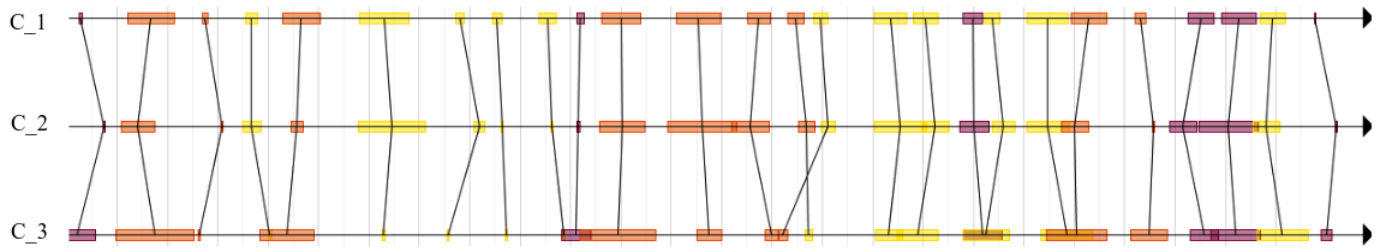
Axe sémantique :

- Ontologique : lien entre signifiés du langage,
  e.g. : « avion » et « véhicule ».
- Lexicale : lien entre signifiants du langage,
  e.g. : « base de données » et « base chimique ».
- Autre : Structures narratives, etc. $\rightarrow$ relations rhétoriques.

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

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.

Alignement :

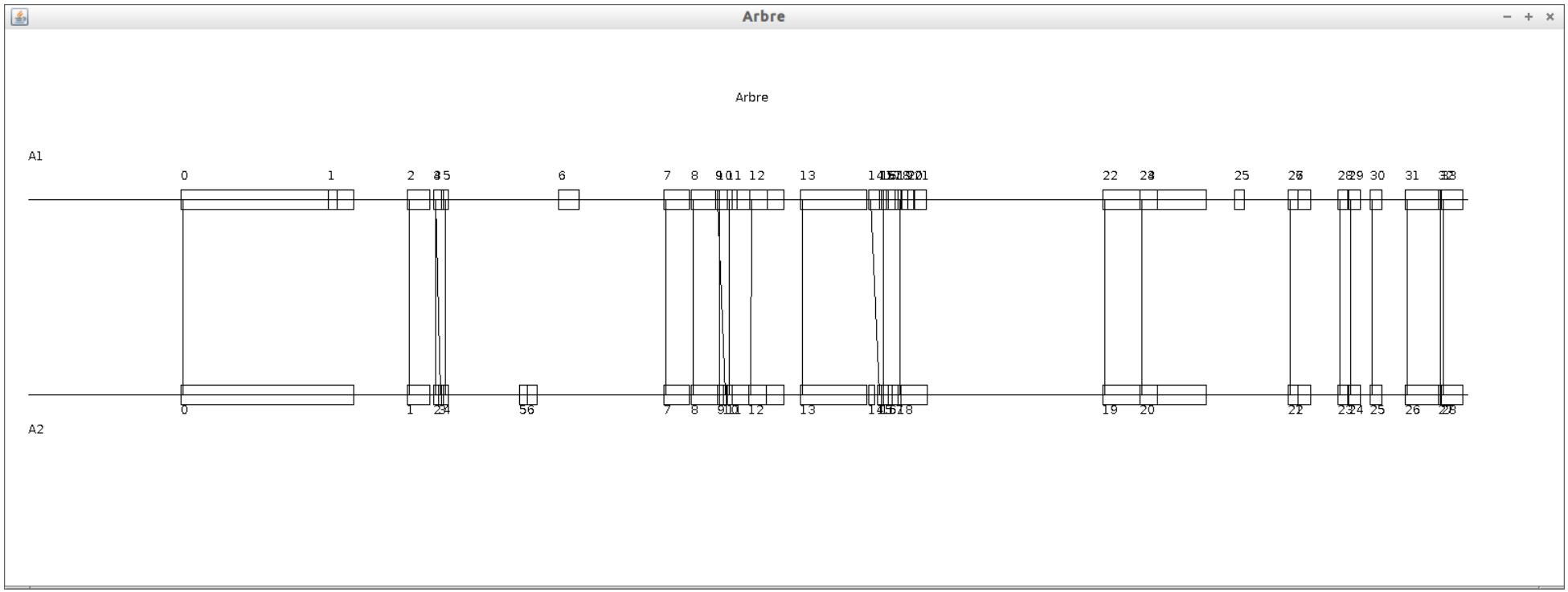
Soit L, $L_{initial}$, L^{-} des listes d'alignements unitaires, $L_{initial}$ est triée, et $\dot{a}$ un alignement candidat [Mathet and Widlöcher, 2011].

algorithm 2 : solution approchée pour l'alignement

1: $L \leftarrow L_{initial}$
2: $i \leftarrow 0$
3: **while** $i < size(L) - 1$:
4: $\quad \dot{a} \leftarrow L[i]$
5: $\quad L^{-} \leftarrow L[i+1, (size(L) - 1)]$
6: $\quad$ Retirer de L^{-} les alignements contenant une unité de $\dot{a}$
7: $\quad L \leftarrow L^{-}$
8: $\quad i \leftarrow i + 1$

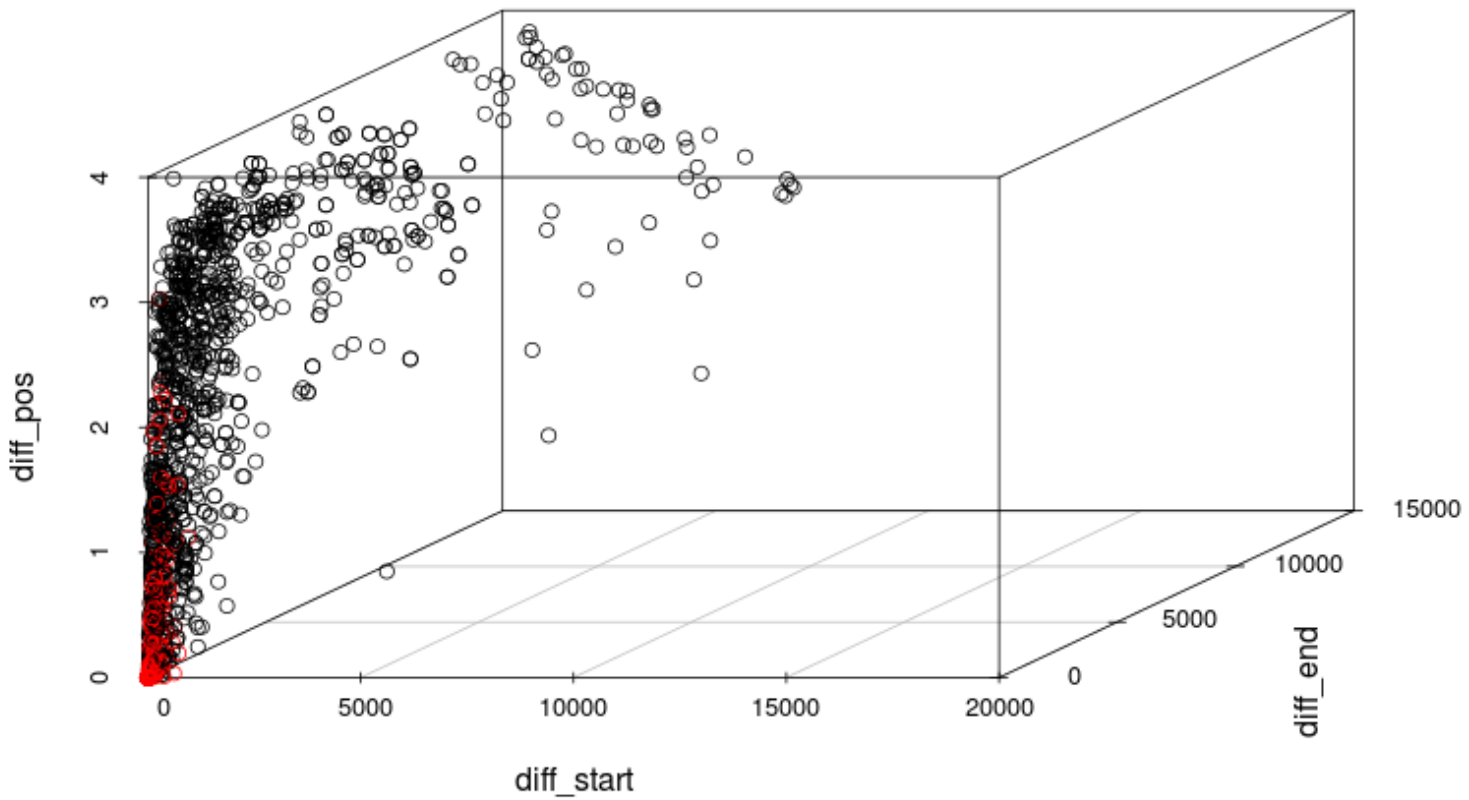
Nettoyage de l'alignement :

- Interface d'édition simple,
- Exemple pour le document « Arbre ».



Alignements unitaires après nettoyage :

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



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	κ	z-score	observations	couverture
isA	**0.45**	18.27	268	36.0%
instanceOf	**0.43**	17.40	196	26.3%
partOf	**0.48**	19.53	39	5.2%
otherOntological	0.28	11.39	42	5.7%
metaLinguistic	0.74	30.40	149	20.0%
other	0.23	09.50	51	6.8%
Corpus	0.49	36.20	745	100%

Références I

Eugene Agichtein and Luis Gravano. Snowball : Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94. ACM, 2000.

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.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction for the web. In *IJCAI*, volume 7, pages 2670–2676, 2007.

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.

Sergey Brin. Extracting patterns and relations from the world wide web. In *The World Wide Web and Databases*, pages 172–183. Springer, 1999.

Références II

Scott Cederberg and Dominic Widdows. Using lsa and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 111–118. Association for Computational Linguistics, 2003.

Philipp Cimiano, Siegfried Handschuh, and Steffen Staab. Towards the self-annotating web. In *Proceedings of the 13th international conference on World Wide Web*, pages 462–471. ACM, 2004.

Philipp Cimiano, Andreas Hotho, and Steffen Staab. Learning concept hierarchies from text corpora using formal concept analysis. *J. Artif. Intell. Res.(JAIR)*, 24 : 305–339, 2005.

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.-P. Fauconnier, M. Kamel, and B. Rothenburger. Une typologie multi-dimensionnelle des structures énumératives pour l'identification des relations termino-ontologiques. In *International Conference on Terminology and Artificial Intelligence (TIA)*, 2013a.

Références III

J.-P. Fauconnier, M. Kamel, B. Rothenburger, and N. Aussenac-Gilles. Apprentissage supervisé pour l'identification de relations sémantiques au sein de structures énumératives parallèles. In *Actes de la 20e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013)*, pages 132–145, 2013b.

J.-P. Fauconnier, L. Sorin, M. Kamel, M. Mojahid, and N. Aussenac-Gilles. Détection automatique de la structure organisationnelle de documents à partir de marqueurs visuels et lexicaux. In *Actes de la 21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, pages 340–351, 2014.

D. Faure and C. Nédellec. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *LREC workshop on adapting lexical and corpus resources to sublanguages and applications*, volume 707, page 30, 1998.

J.L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5) :378, 1971.

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 IV

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 enumératives pour la construction d'ontologie. In *Journées Francophones d'Ingénierie des Connaissances (IC 2011)*, pages 505–522, Annecy, 2011.

M. Kamel, B. Rothenburger, and J.-P. Fauconnier. Identification de relations sémantiques portées par les structures énumératives paradigmatiques. *Revue d'Intelligence Artificielle*, Ingénierie des Connaissances, 2014.

Alessandro Lenci and Giulia Benotto. Identifying hypernyms in distributional semantic spaces. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1 : Proceedings of the main conference and the shared task, and Volume 2 : Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 75–79. Association for Computational Linguistics, 2012.

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.

Références V

D. Marcu. Automatic discourse parsing. In K. Brown, editor, *Encyclopedia of Language and Linguistics*. Elsevier, 2nd edition, 2006.

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.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

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

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.

D. Sanchez and A. Moreno. Web-scale taxonomy learning. In *Proceedings of Workshop on Extending and Learning Lexical Ontologies using Machine Learning (ICML 2005)*, pages 53–60, Bonn, Germany, 2005.

Références VI

R. Snow, D. Jurafsky, and A. Y. Ng. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems*, volume 17, 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.