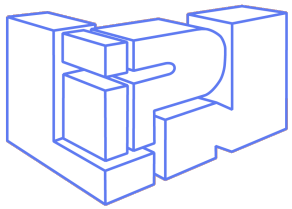
Construction d'ontologies à partir de textes
L'apport de l'analyse de concepts formels

Thibault Mondary

Laboratoire d'Informatique de Paris-Nord



27 mai 2011

Contexte

Pourquoi partir de textes ?

- plus accessibles que les experts
- connaissances stabilisées et partagées sur un domaine
- facilitent le retour au texte

« Une ontologie est une spécification formelle d'une conceptualisation d'un ***domaine****, partagée par un groupe de personnes, qui est établie selon un certain point de vue imposé par l'application construite. »*
[Studer et al., 1998]

Conceptualisation

Comment identifier et mettre en relation les unités représentatives du domaine ?

- Une étape **difficile**
  - grande quantité d'informations à appréhender
  - connaissances du domaine souvent implicites
  - prisme de la langue (ambiguïté, synonymie, ellipse)
  - expertise nécessaire
  - choix de modélisation
- Une étape qui *ne peut pas être totalement automatisée*
- Une étape à outiller

Problématique

Outiller la construction d'ontologies à partir de textes (COT)

- Repérage des éléments clés : analyse terminologique
- Apport de la structure : analyse de concepts formels (ACF)

Comment combiner ACF et approche terminologique ?

Sommaire

Sommaire

Quelques définitions informelles

Composants d'une ontologie

- Concepts
- Relations hiérarchiques
- Instances de concepts
- Relations spécialisées, instances de relations, axiomes, règles

Un concept

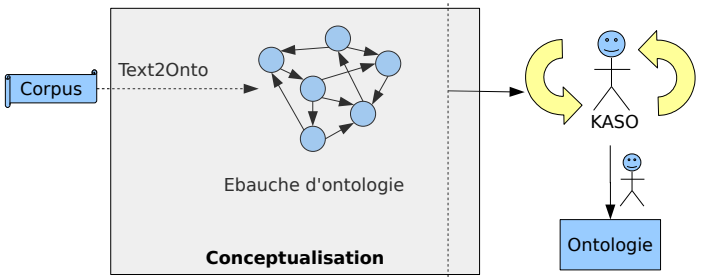
Une intension : ce qui le caractérise

Une extension : ce qu'il regroupe

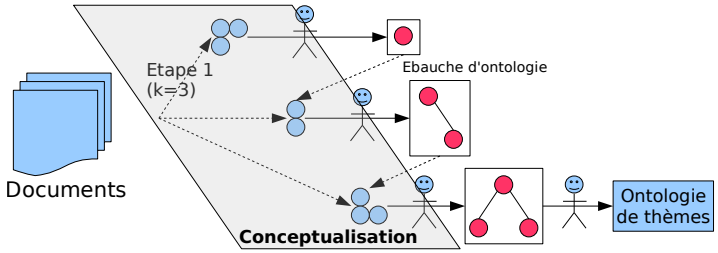
Une dénotation : comment le nommer ?

Trois approches de la COT

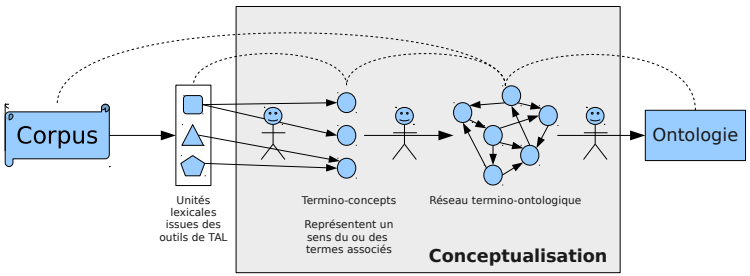
Approche automatique : Text2Onto [Cimiano & Völker, 2005]



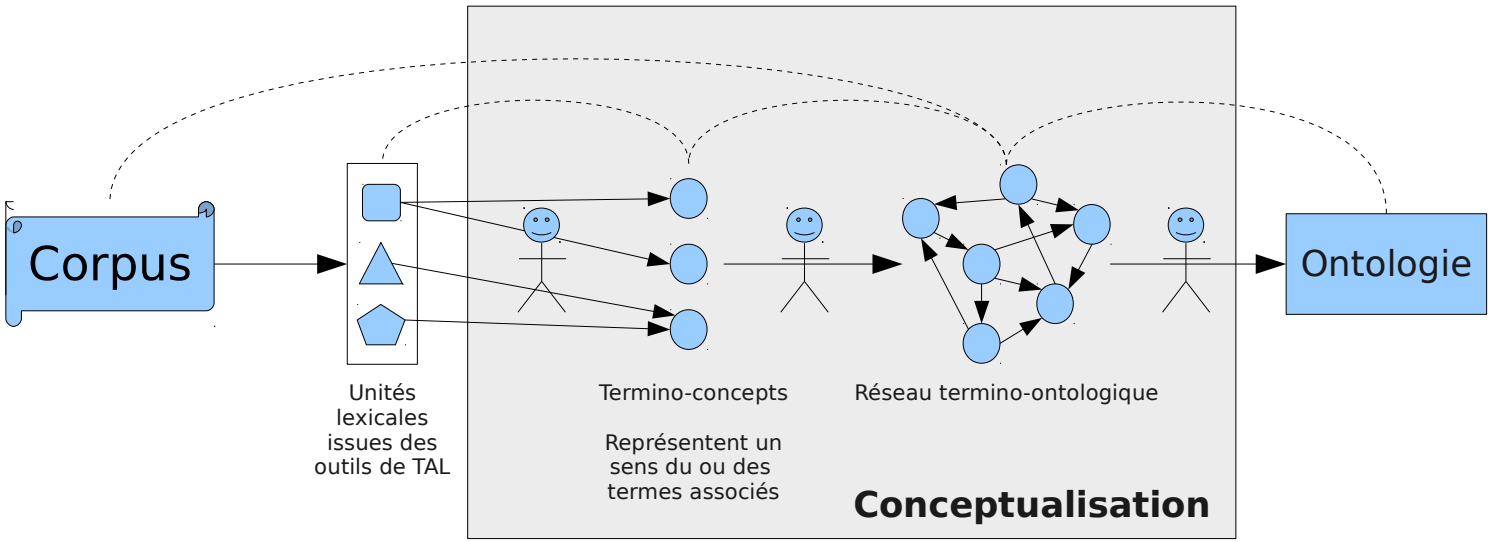
Approche itérative assistée : Ontogen [Fortuna et al., 2006]



Approche terminologique libre assistée : Terminae



Approche terminologique libre assistée : Terminae



- Une méthodologie complète : du texte à l'ontologie
- Une séparation claire des niveaux textuel et conceptuel
- Un lien avec le texte conservé

Le réseau termino-ontologique doit être structuré à la main

Structurer une liste d'éléments textuels

Approches directes

- Repérage de l'hyperonymie [Hearst, 1992]
- Repérage de relations spécialisées
  [Morin, 1999, Séguéla & Aussenac-Gilles, 1999]
- Résultats généralement fiables
- Dépend de la nature du corpus, de la langue et du domaine

Approches indirectes

Hypothèse distributionnelle [Harris et al., 1989]

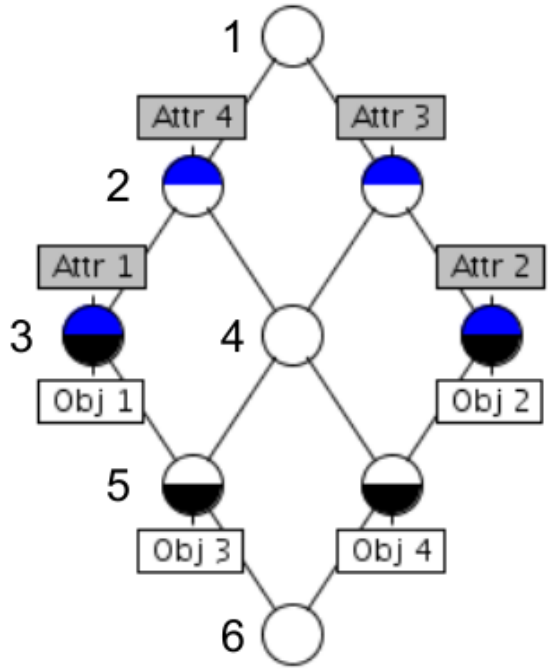
- Par similarité [Hindle, 1990]
  - Mesure de distance entre éléments
  - Mots plus que termes
- Booléennes [Ganter & Wille, 1999, Cimiano et al., 2005]
  - Approche ensembliste
  - Termes et propriétés facilement pris en compte

Analyse de concepts formels (ACF), principes

- Méthode de regroupement symbolique proposée par [Ganter & Wille, 1999]
- Découverte de tous les regroupements (**concepts formels**) possibles d'éléments (**objets formels**) avec des propriétés (**attributs formels**) en commun
- Les objets et les attributs sont organisés dans un **contexte formel**, triplet $\mathbb{K} = (G, M, I)$
- Les concepts formels :
  - possèdent une intension et une extension
  - forment une structure de treillis

ACF, exemple

	Attr 1	Attr 2	Attr 3	Attr 4
Obj 1	×			×
Obj 2		×	×	
Obj 3	×		×	×
Obj 4		×	×	×



1. TOP, « attributs formels partagés par tous les objets »
2. Concept formel $(\{Obj1, Obj3, Obj4\}, \{Attr4\})$ sans extension propre
3. Concept formel $(\{Obj1, Obj3\}, \{Attr1, Attr4\})$
4. Concept formel sans intension ni extension propre (« blanc »)
5. Concept formel sans intension propre
6. BOTTOM, « objets formels qui possèdent tous les attributs »

ACF et textes : le contexte formel

Principes

- Utilisation de la structure syntaxique des phrases
- Prise en compte des relations sujet/verbe et cod/verbe :
  - Les objets formels sont les sujets et les cod des phrases
  - Les attributs formels sont les verbes associés, suffixés de `-ABLE` en cas de cod

Exemple

« *Une grève paralyse l'entreprise. Les travailleurs refusent le travail dominical. Les dirigeants refusent la grève.* »

	paralyser	paralyser-ABLE	refuser	refuser-ABLE
grève	×			×
entreprise		×		
travailleur			×	
dirigeant			×	
travail				×

ACF et textes : le contexte formel

Principes

- Utilisation de la structure syntaxique des phrases
- Prise en compte des relations sujet/verbe et cod/verbe :
  - Les objets formels sont les sujets et les cod des phrases
  - Les attributs formels sont les verbes associés, suffixés de `-ABLE` en cas de cod

Exemple

« *Une **grève** **paralyse** l'entreprise. Les travailleurs refusent le travail dominical. Les dirigeants **refusent** la **grève**.* »

	paralyser	paralyser-ABLE	refuser	**refuser-ABLE**
grève	×			×
entreprise		×		
travailleur			×	
dirigeant			×	
travail				×

Qui utilise l'ACF ?

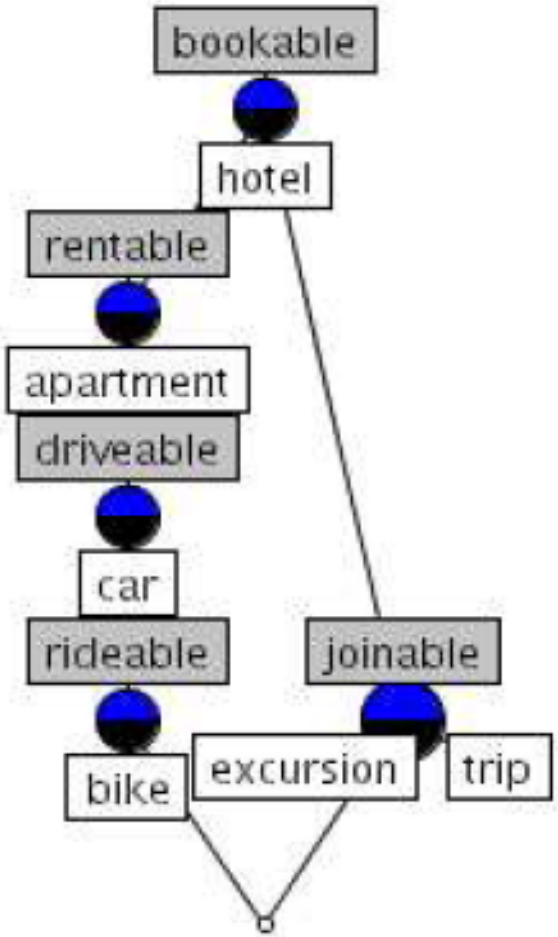
[Cimiano et al., 2005]

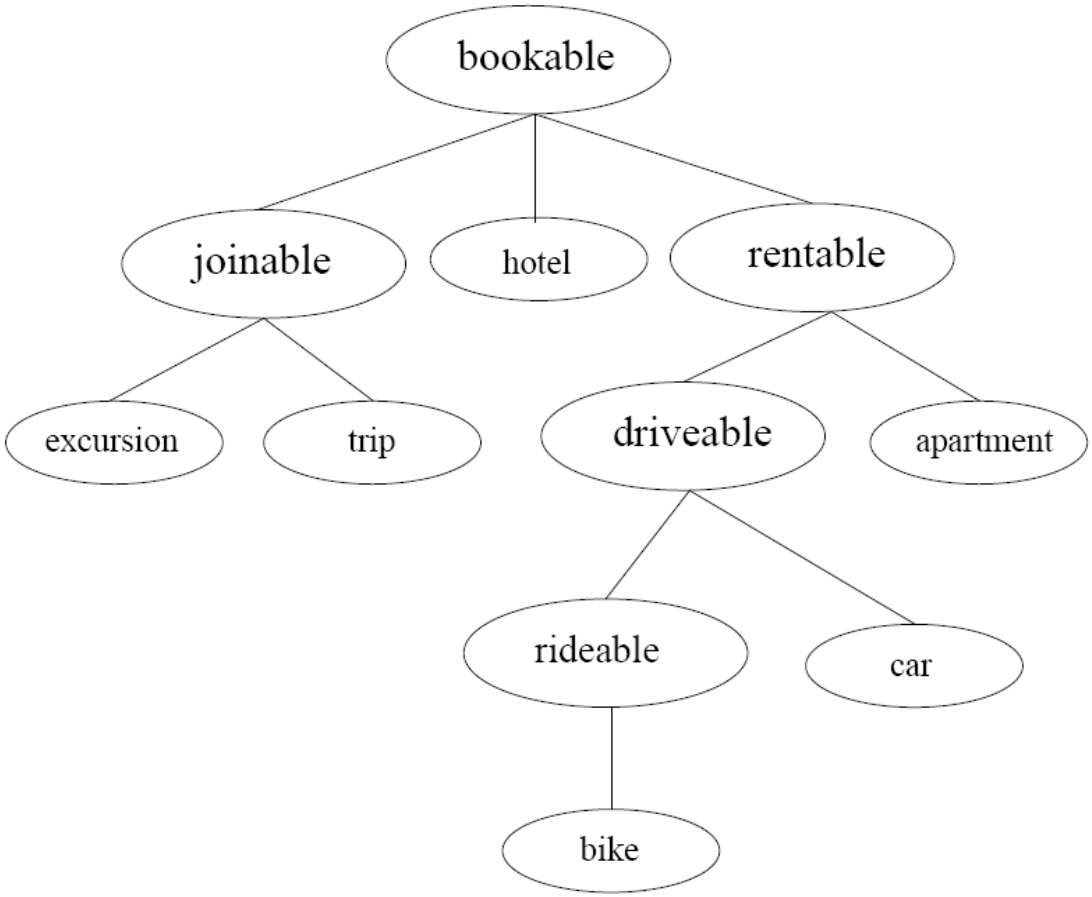
- Compare l'ACF avec les classifications hiérarchiques (ACF meilleur)
- Ressource de départ : corpus de grande taille
- Objets formels : mots
- Contexte : relations syntaxiques
- Traduction en ontologie : automatique

[Bendaoud et al., 2008]

- Enrichissement d'ontologies
- Ressource de départ : ontologie + corpus
- Objets formels : instances de l'ontologie initiale
- Contexte : relations syntaxiques

ACF => ontologie pour [Cimiano et al., 2005]





Bilan

- Approche terminologique pertinente mais largement manuelle
- ACF doit pouvoir se combiner naturellement avec les termes

Hypothèse de travail

La langue n'est pas une collection d'objets et d'attributs formels

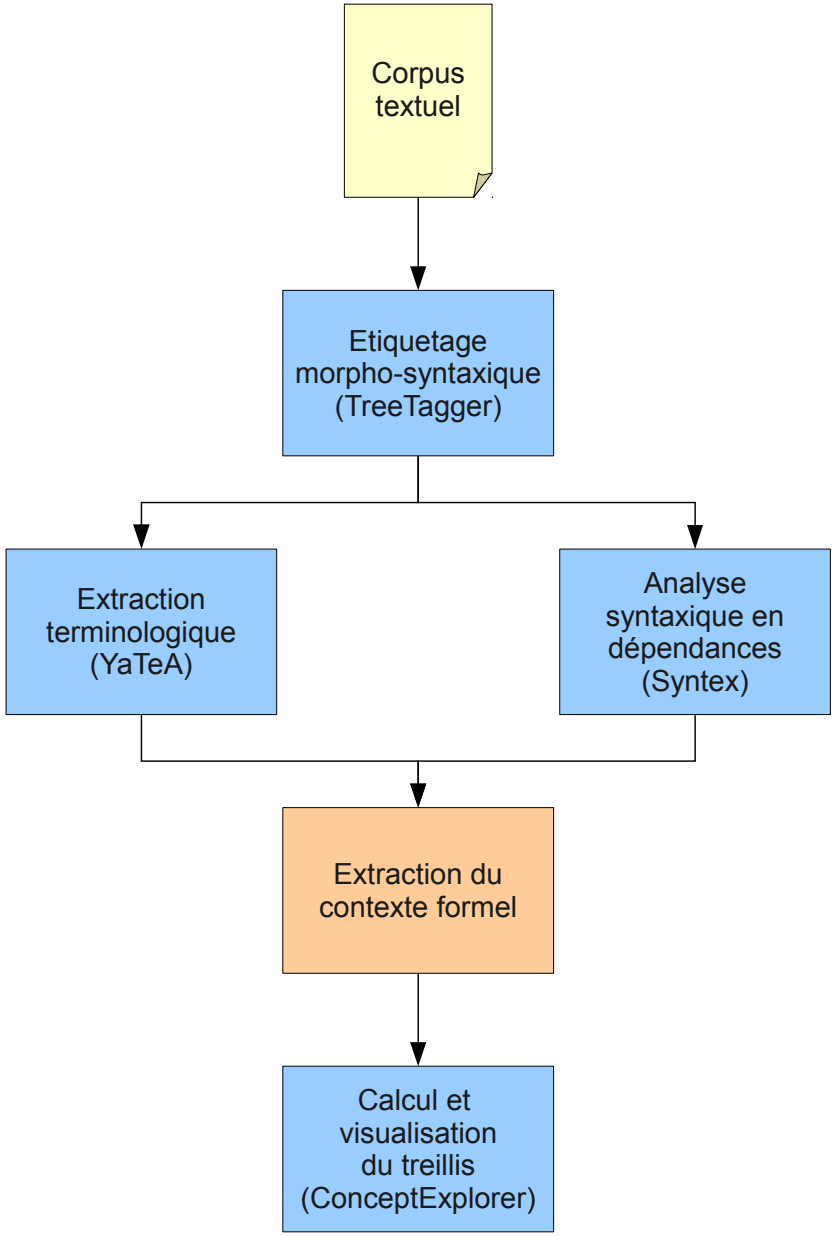
- Les mots / termes : concepts, sous-concepts, instances
- Les prédicats linguistiques : propriétés essentielles ou accessoires

Problématique

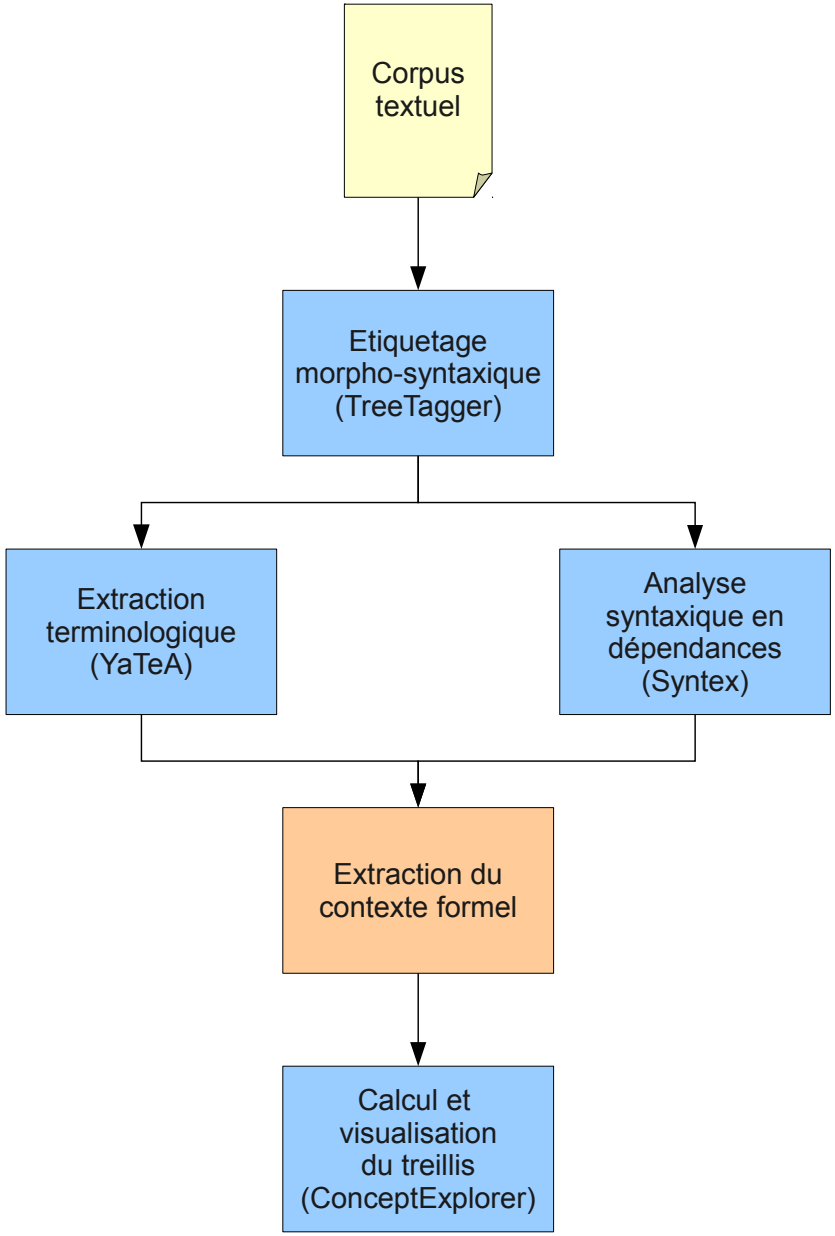
- Que peut-on attendre de l'application de l'ACF sur les textes ?
- Quelles précautions prendre ?

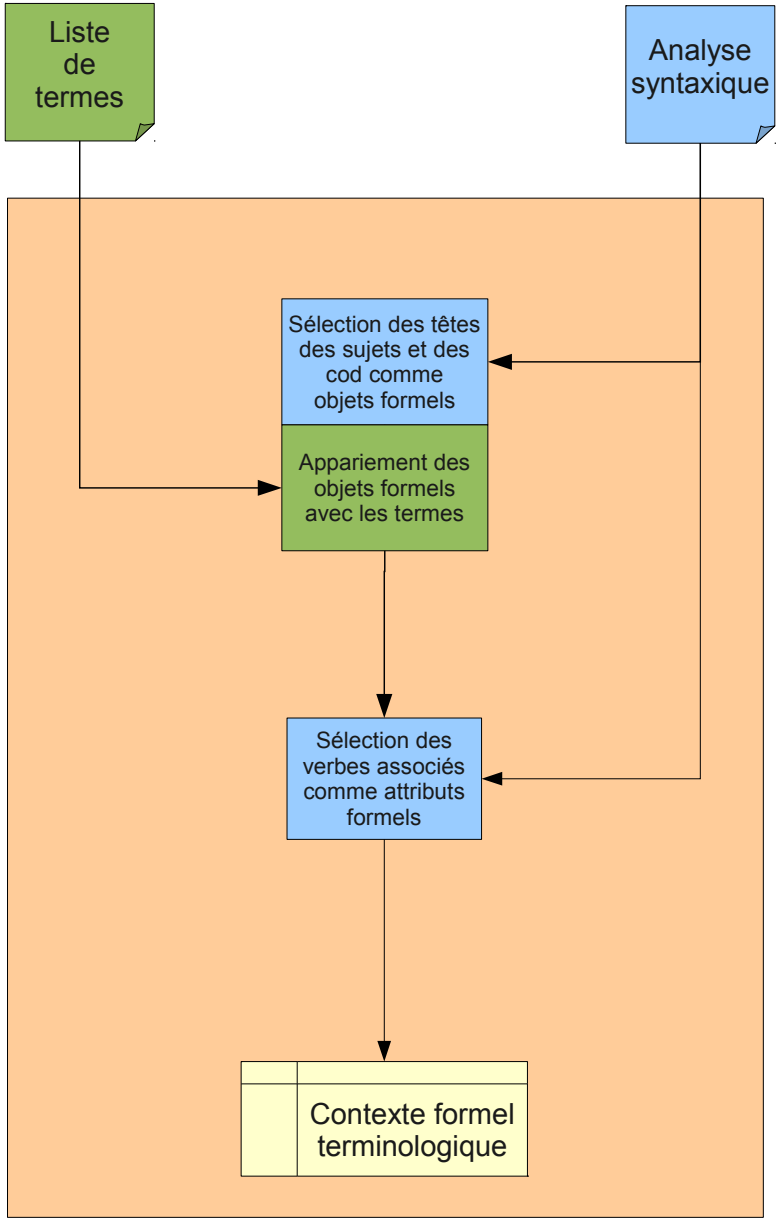
Sommaire

Méthodologie



Méthodologie





Exploitation de l'analyse syntaxique

Objectifs

- Identifier des relations prédicats / arguments
- Passer d'une analyse de surface à une analyse syntaxico-sémantique profonde

Cas simple

« une grève paralyse l'entreprise »

- sujet(paralyser, grève)
- cod(paralyser, entreprise)

Cas complexe

« L'étoile est bleue »

- sujet(être, étoile) —> sujet(être bleue, étoile)

Exploitation de l'analyse syntaxique (2)

Démarche

- Utilisation des sorties de Syntex
- Interprétation en analyse profonde à l'aide d'heuristiques

Tournures prises en compte

Attribut du sujet : « Jean **est heureux** »

Coordination : « J'aime le pain **et** le chocolat »

Anaphore relative : « Le voisin **qui** parle anglais »

Auxiliaire : « Jean **a regardé** Marie »

Tournures non prises en compte

Autre type d'anaphore : « **Elle** explose en supernova »

Modalité : « L'étoile **peut** exploser »

Expériences : corpus

Trois corpus différents mais de taille comparable, en français :

- le droit international du travail (14 conventions) : 53 321 mots
- la cuisine (392 recettes) : 54 227 mots
- l'astronomie (27 pages Wikipédia) : 54 121 mots

Expériences : corpus (2)

Un corpus difficile : le droit

« *La présente convention s'applique à tout navire de mer, de propriété publique ou privée, qui est immatriculé dans le territoire de tout Membre pour lequel la convention est en vigueur et qui est normalement affecté à la navigation maritime commerciale.* »

Une langue simplifiée : la cuisine

« *Dans un bol, mettre l'eau et le sucre semoule, ajouter le kirsch. Il faut attendre que la totalité du sucre soit bien dissoute. Choisir un moule à charlotte ou à flan.* »

Des textes explicites : l'astronomie

« *En astronomie, un amas globulaire est un amas stellaire très dense, contenant typiquement une centaine de milliers d'étoiles distribuées dans une sphère dont la taille varie de 20 à quelques centaines d'années-lumière. Les étoiles de ces amas sont généralement des géantes rouges.* »

Des résultats hétérogènes selon les corpus

Sans les termes	Astronomie	Droit	Cuisine
Objets formels	774	432	456
Attributs formels	1089	416	463
Concepts formels	1555	514	1777
Hauteur du treillis	9	6	11
Nombre d'arêtes	4170	1179	5371

Avec les termes	Astronomie	Droit	Cuisine
Candidats-termes	4881	3124	3607
Objets formels	1526 (+752)	770 (+338)	1010 (+554)
Attributs formels	1089	416	463
Concepts formels	1351 (-204)	513 (-1)	1252 (-525)
Hauteur du treillis	8 (-1)	5 (-1)	9 (-2)
Nombre d'arêtes	3241 (-929)	1112 (-67)	3460 (-1911)

Des résultats hétérogènes selon les corpus

Sans les termes	Astronomie	Droit	Cuisine
Objets formels	**774**	**432**	**456**
Attributs formels	1089	416	463
Concepts formels	1555	514	1777
Hauteur du treillis	9	6	11
Nombre d'arêtes	4170	1179	5371

Avec les termes	Astronomie	Droit	Cuisine
Candidats-termes	4881	3124	3607
Objets formels	**1526 (+752)**	**770 (+338)**	**1010 (+554)**
Attributs formels	1089	416	463
Concepts formels	1351 (-204)	513 (-1)	1252 (-525)
Hauteur du treillis	8 (-1)	5 (-1)	9 (-2)
Nombre d'arêtes	3241 (-929)	1112 (-67)	3460 (-1911)

Des résultats hétérogènes selon les corpus

Sans les termes	Astronomie	Droit	Cuisine
Objets formels	774	432	456
Attributs formels	**1089**	**416**	**463**
Concepts formels	1555	514	1777
Hauteur du treillis	9	6	11
Nombre d'arêtes	4170	1179	5371

Avec les termes	Astronomie	Droit	Cuisine
Candidats-termes	4881	3124	3607
Objets formels	1526 (+752)	770 (+338)	1010 (+554)
Attributs formels	**1089**	**416**	**463**
Concepts formels	1351 (-204)	513 (-1)	1252 (-525)
Hauteur du treillis	8 (-1)	5 (-1)	9 (-2)
Nombre d'arêtes	3241 (-929)	1112 (-67)	3460 (-1911)

Des résultats hétérogènes selon les corpus

Sans les termes	Astronomie	Droit	Cuisine
Objets formels	774	432	456
Attributs formels	1089	416	463
Concepts formels	**1555**	**514**	**1777**
Hauteur du treillis	**9**	**6**	**11**
Nombre d'arêtes	**4170**	**1179**	**5371**

Avec les termes	Astronomie	Droit	Cuisine
Candidats-termes	4881	3124	3607
Objets formels	1526 (+752)	770 (+338)	1010 (+554)
Attributs formels	1089	416	463
Concepts formels	**1351 (-204)**	**513 (-1)**	**1252 (-525)**
Hauteur du treillis	**8 (-1)**	**5 (-1)**	**9 (-2)**
Nombre d'arêtes	**3241 (-929)**	**1112 (-67)**	**3460 (-1911)**

Interprétation du treillis en ontologie

Objectifs

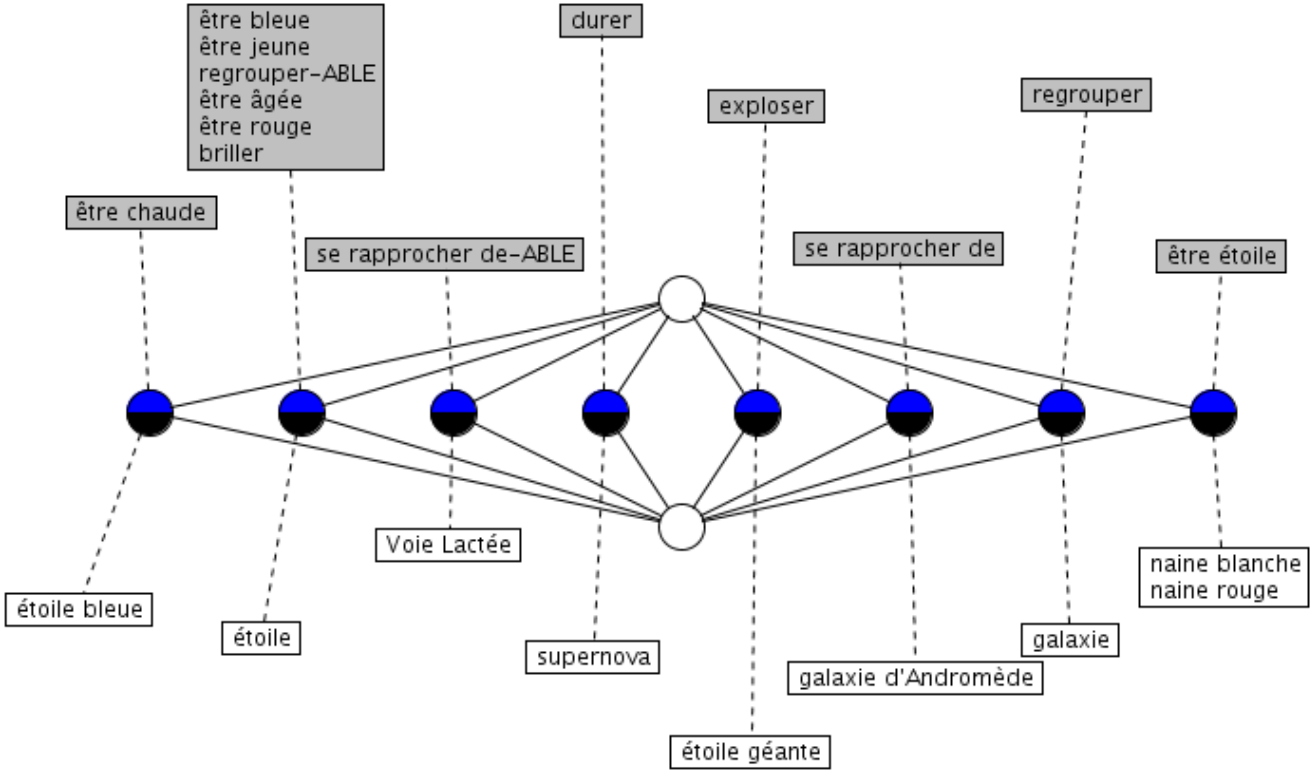
- Dénotation des concepts
  - l'intension
  - l'extension
  - autre chose ?
- Structure du treillis
  - valider les liens hiérarchiques
  - valider la granularité de la description

Exemple

Exemple jouet sur le domaine de l'astronomie

Une étoile bleue est très chaude.
Les étoiles brillent.
Les naines rouges sont des étoiles.
Une galaxie regroupe des étoiles.
Cette étoile est rouge et âgée.
Cette étoile est bleue et jeune.
Une naine blanche était autrefois une étoile.
La galaxie d'Andromède se rapproche de la Voie Lactée.
Des étoiles géantes explosent parfois en supernova.
Une supernova ne dure pas très longtemps.

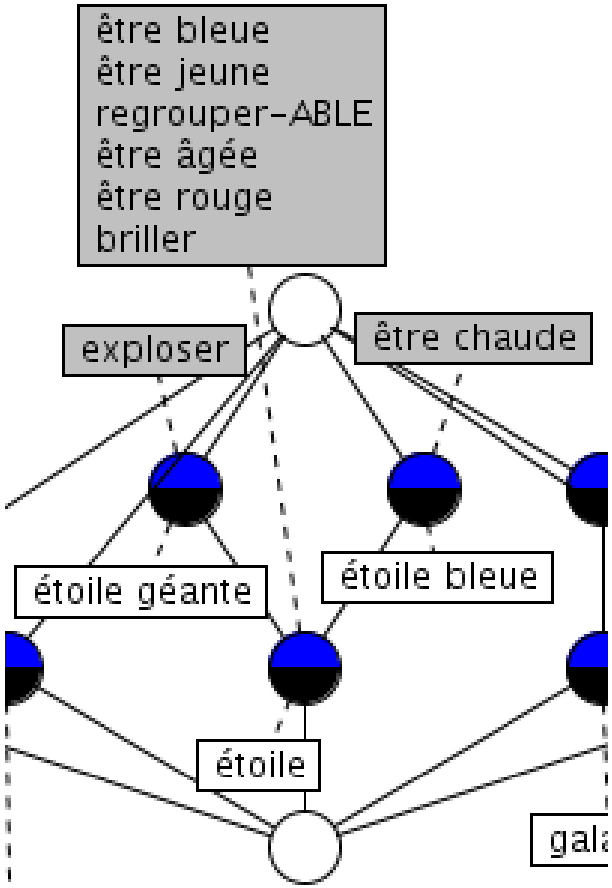
Interprétation du treillis : illustration



Problèmes

- Pas de structure
- Des regroupements d'objets erronés (extension)
  « Une naine blanche était autrefois une étoile »
- Des conjonctions d'attributs incohérentes (intension)

Interprétation du treillis : illustration (2)



Subsomption inversée autour de `étoile`

L'objet formel dénote

- **un concept** (les étoiles brillent)
- **un sous-concept** (une étoile bleue est très chaude)
- **une instance** (cette étoile est rouge)

Les attributs formels dénotent des propriétés

- **essentielles** (briller)
- **accessoires** (être bleue)

Bilan

Des problèmes

- Des résultats hétérogènes sur des corpus de taille comparable
  - Nombre important de concepts formels pour de petits corpus
  - Une variation du simple au double de leur nombre
- Une interprétation difficile

Sommaire

Prévoir le nombre de concepts formels

Problématique

- L'ACF fournit des résultats variables sur des corpus de taille similaire
- Est-il possible d'estimer le nombre de concepts formels à l'aide d'indicateurs de surface ?

Indicateurs étudiés

1. Au niveau du vocabulaire
2. Au niveau des phrases

Des indicateurs au niveau du vocabulaire ?

	Droit	Astronomie	Cuisine
Nombre de concepts formels	**513**	**1351**	**1252**
Taille du vocabulaire lemmatisé	2415	3978	2114
Taille du vocabulaire fléchi	3599	5827	3043

Pas de relation...

Des indicateurs au niveau des phrases ?

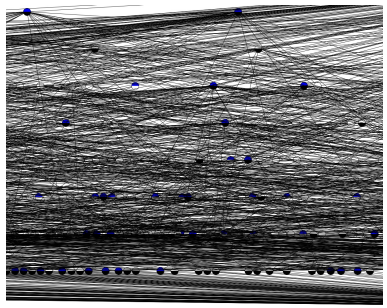
Idée : plus les phrases sont complexes, moins on retrouve de structures prédicatives exploitables

	Droit	Astronomie	Cuisine
Nombre de concepts formels	**513**	**1351**	**1252**
Nombre de mots	53321	54121	54527
Nombre de phrases	970	2287	4074
Phrase la plus longue (en mots)	824	200	78
Nombre moyen de mots par phrase	**55**	**23**	**13**

Estimer le nombre de concepts formels : bilan

- Indicateurs de surface élémentaires : non probant
- Combinaisons d'indicateurs : une piste à explorer

Maîtriser le nombre de concepts formels



Problème

- L'ACF est très prolixe en concepts formels => difficile à afficher et à s'approprier
- Il faut pouvoir contrôler le nombre de concepts formels

Pistes pour contrôler le nombre de concepts formels

1. Lors de la constitution du corpus
2. Au moment de la construction du contexte formel
3. Une fois le contexte formel construit
4. Au moment de la visualisation

Filtrage lors de la construction du contexte formel

Idée : diminuer le nombre de concepts formels en jouant sur la prise en compte des termes

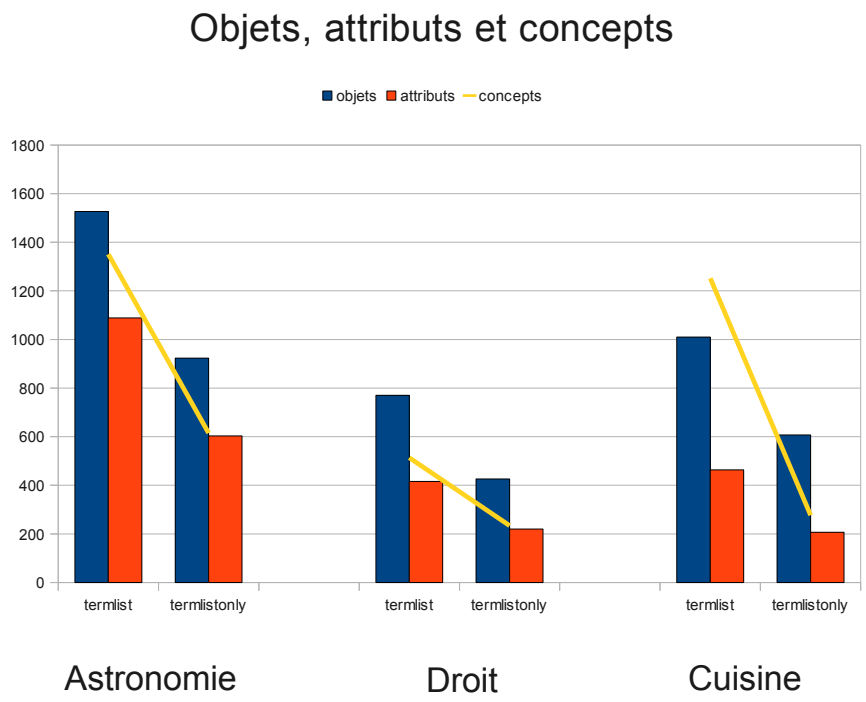
Trois variantes de construction du contexte formel

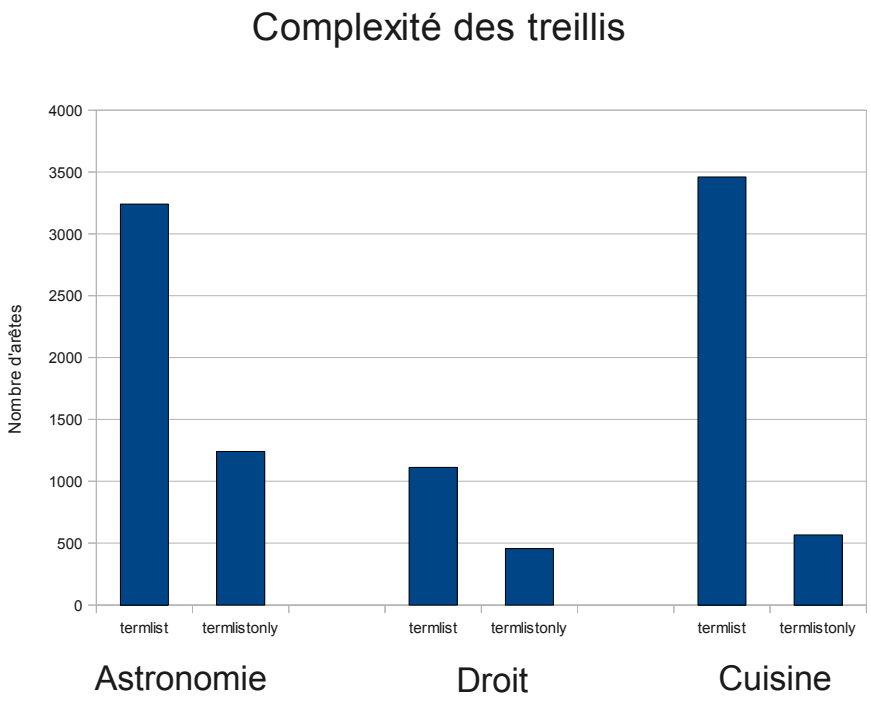
usuelle : Objets formels = mots

termlist : Objets formels = termes de préférence ou mots

termlistonly : Objets formels = termes

Filtrage lors de la construction du contexte formel (2)





Filtrage une fois le contexte formel construit

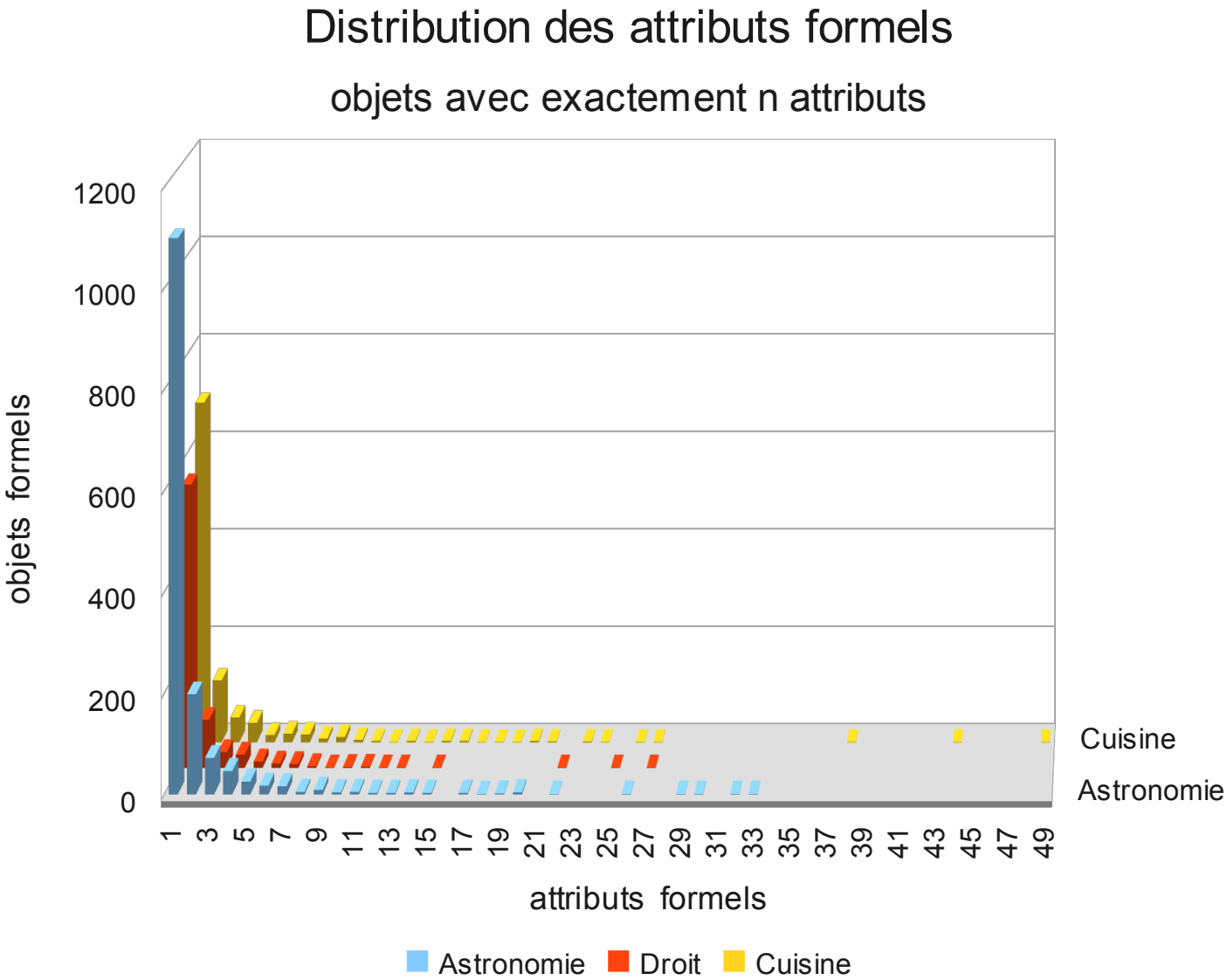
Idée : diminuer le nombre de concepts formels en jouant sur la relation objet / attribut

Deux approches de filtrage

1. Les objets formels qui possèdent n attributs
2. Les attributs formels qui s'appliquent à n objets

Filtrage une fois le contexte formel construit (2)

Idée : Les objets qui possèdent trop ou trop peu d'attributs sont moins utiles



Filtrage une fois le contexte formel construit (3)

Idée : Les attributs formels qui apparaissent trop ou trop peu sont moins utiles

Attributs apparaissant avec un seul objet

	termlist	termlistonly
Astronomie	576 (53 %)	355 (59 %)
Droit	155 (37,2 %)	97 (44 %)
Cuisine	192 (41,5 %)	109 (52,6 %)

Les cinq attributs les plus dispersés

Astronomie : posséder-ABLE, permettre, former-ABLE, posséder, devoir

Droit : devoir, porter-ABLE, ratifier-ABLE, enregistrer-ABLE, assurer-ABLE

Cuisine : ajouter-ABLE, verser-ABLE, mélanger-ABLE, retirer-ABLE, mettre-ABLE

Maîtriser le nombre de concepts formels : bilan

Incorporer les termes uniquement

+ Focalise sur les éléments pertinents
= Diminue les regroupements (cf. *subsomption directe*)

Supprimer les attributs les plus dispersés

+ Élimine la phraséologie inutile (*cf. droit*)
– Risque de supprimer des connaissances utiles (*cf. cuisine*)

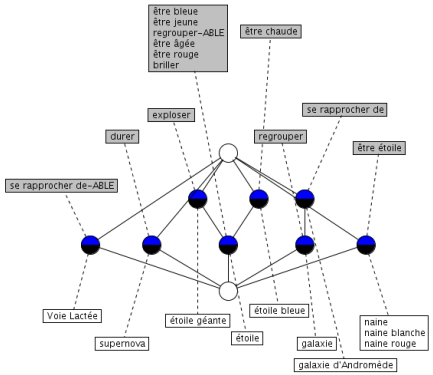
Supprimer les attributs non dispersés

– Aucune influence sur les regroupements
– Diminue l'interprétabilité (supprime des intensions propres)

Supprimer les objets qui possèdent trop ou trop peu d'attributs

– Aucun contrôle sur les regroupements supprimés
– Diminue l'interprétabilité (supprime des extensions)

Produire un treillis interprétable



Objectifs

1. proposer des attributs plus interprétables
2. prendre en compte la négation
3. focaliser l'analyse sur une famille de termes
4. injecter de la connaissance dans le treillis
5. corriger le treillis (subsomptions inversées)

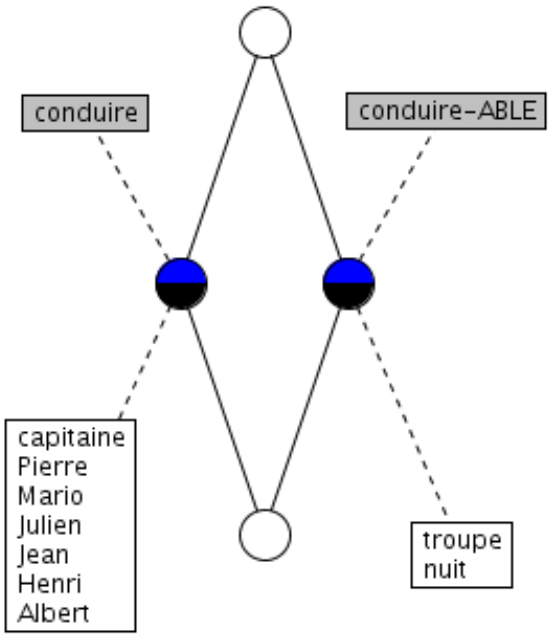
Des attributs plus interprétables

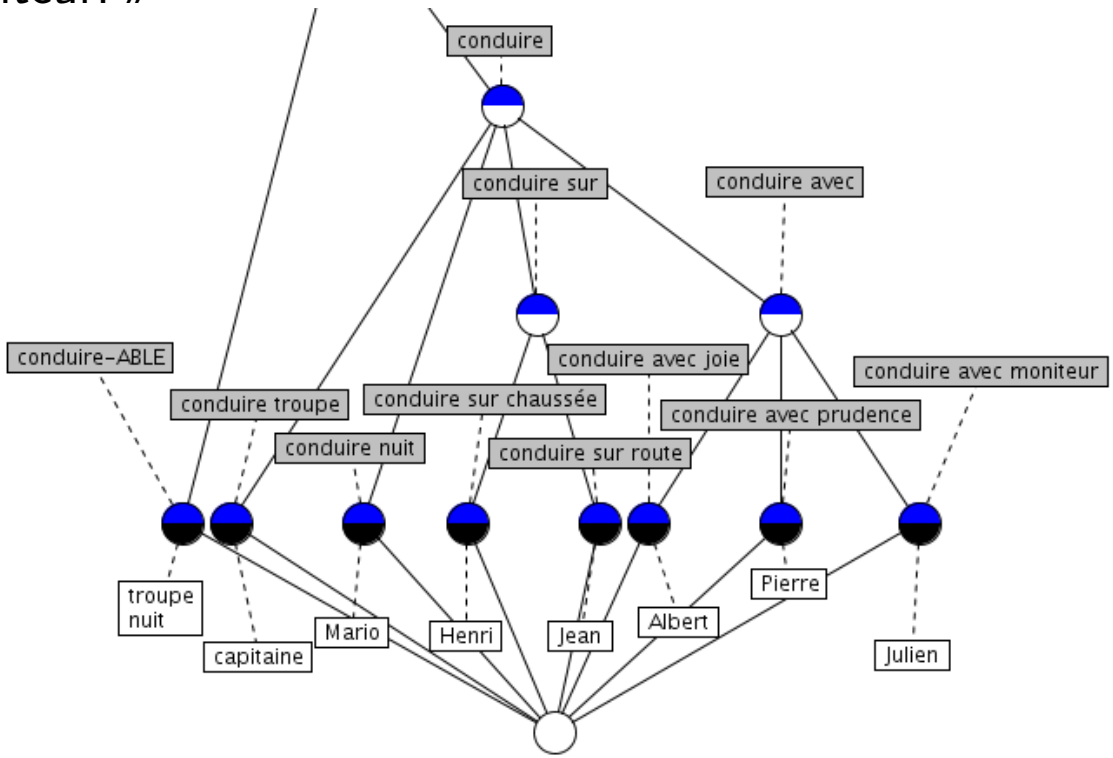
Constatation : les verbes seuls ne sont pas assez informatifs

Idée : prendre en compte le groupe verbal

Attributs formels étendus : verbes + complément prépositionnel ou COD

Exemple : « Jean conduit sur la route. Pierre conduit avec prudence. Henri conduit sur la chaussée. Albert conduit avec joie. Le capitaine conduit ses troupes. Mario conduit la nuit. Julien conduit avec son moniteur. »





Prendre en compte la négation

« Les astéroïdes **ne sont pas** les satellites d'une planète »

- La négation devrait être prise en compte dans le treillis :
  - éviter les regroupements faux
  - éviter d'induire l'ontologue en erreur

Prise en compte de la négation

- Recherche d'un marqueur de négation relié au verbe
- Marquage (NEG) ou suppression de l'attribut

Le repérage de la négation n'est pas trivial :
« En astronomie, les naines rouges sont les étoiles les moins massives ; en deçà, ce sont les naine brunes, qui ne sont pas vraiment des étoiles »

Focaliser l'analyse sur une famille de termes

Idée : Construire un micro-treillis autour d'un terme pertinent du domaine

Objet formel étendu : objet formel qui est composé d'un syntagme nominal maximal relativement au corpus. (ex. trou noir, trou noir supermassif, trou noir galactique)

Démarche : construire un treillis avec le terme et les objets formels étendus associés

Avantages

- permet de se focaliser rapidement sur un terme
- permet d'explorer les notions connexes (ex. trou noir supermassif constitue le centre d'une galaxie)

Injecter de la connaissance dans le treillis

Objectifs

- injecter dans le treillis des relations de subsomption directes
  - la structure syntaxique des termes
  - les patrons d'hyperonymie
- Attention au risque de subsomption inversée !

Idée : raisonner en termes de propriétés plutôt que d'emplois

Initial	Attr 1	Attr 2	Attr 3	Attr 4
Moteur				×
Moteur diesel	×	×		
Moteur essence		×	×	

Emplois	Attr 1	Attr 2	Attr 3	Attr 4
Moteur	×	×	×	×
Moteur diesel	×	×		×
Moteur essence		×	×	×

Propriétés	Attr 1	Attr 2	Attr 3	Attr 4
Moteur				×
Moteur diesel	×	×		×
Moteur essence		×	×	×

Corriger le treillis

Rappel du problème de la subsomption inversée

- on observe des mots et leurs emplois
- on veut des objets et leurs attributs
- cette transformation n'est pas triviale
  - Les termes renvoient tantôt à des concepts, des sous-concepts ou des instances
  - Les propriétés sont tantôt essentielles (conjonction des attributs *valide*), tantôt accessoires (conjonction *invalide*)

Corriger le treillis (2)

Définitions

Attribut générique : traduit une propriété du concept évoqué par l'objet formel (ex. comète :se composer de trois parties)

Attribut spécifique : traduit une propriété d'un sous-concept ou d'une instance du concept évoqué par l'objet formel (ex. comète :percuter Jupiter)

Des difficultés de repérage

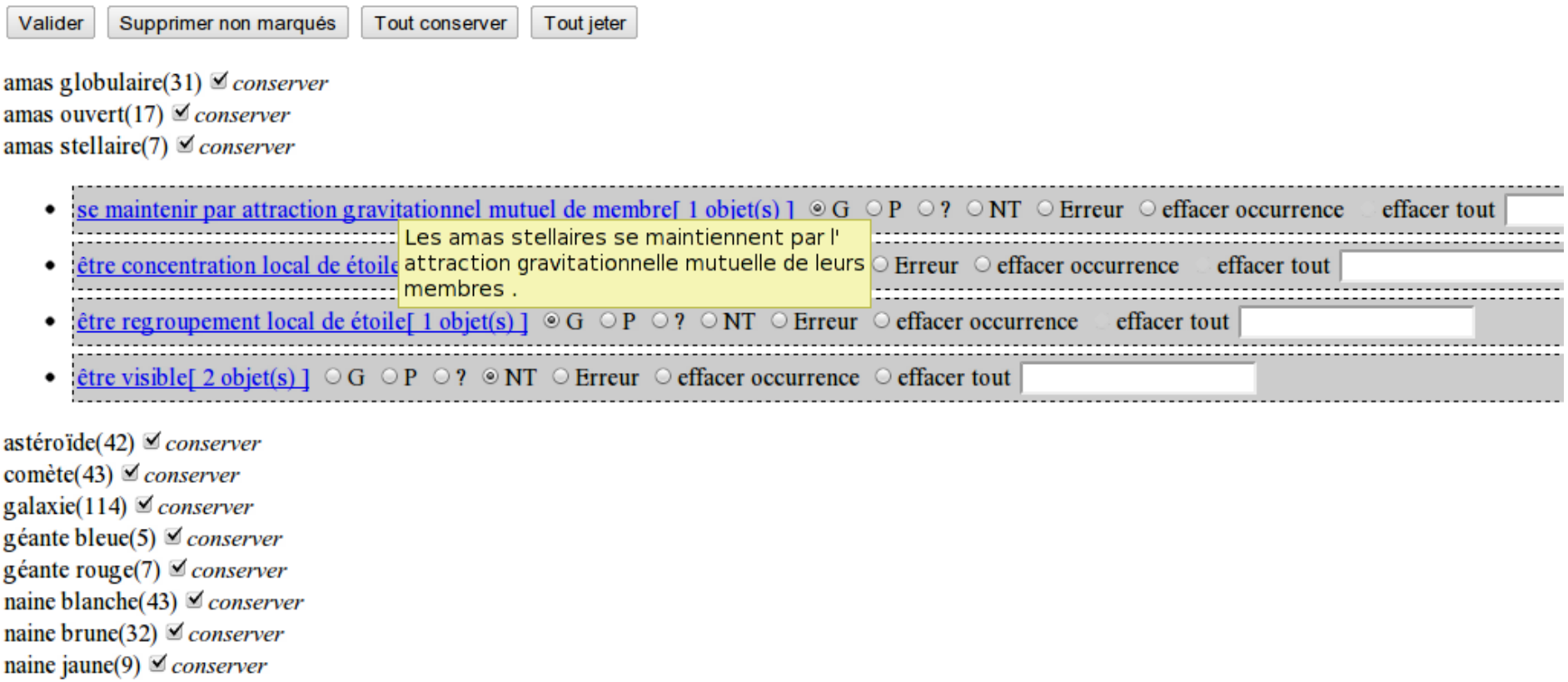
- des connaissance du domaine
- un retour au texte (marqueurs de spécificité ou de généricité)

Pourquoi est-ce important ?

- ACF : conjonction des attributs
- Les attributs spécifiques : incohérences + subsomption inversée
- Risque moindre pour les entités nommées dont l'interprétation en objet est plus naturelle

Corriger le treillis (3)

Idée : séparer les attributs génériques des spécifiques

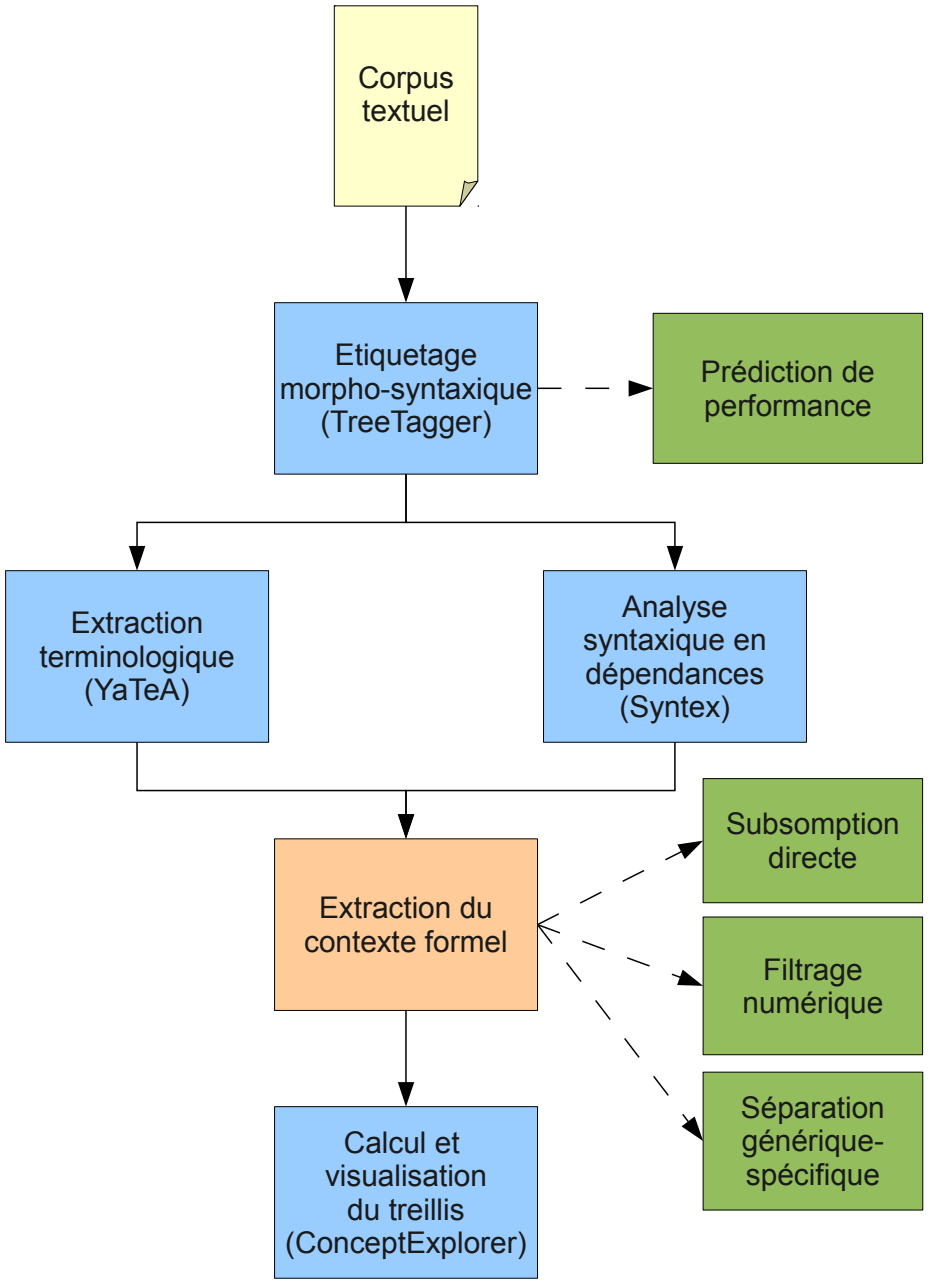


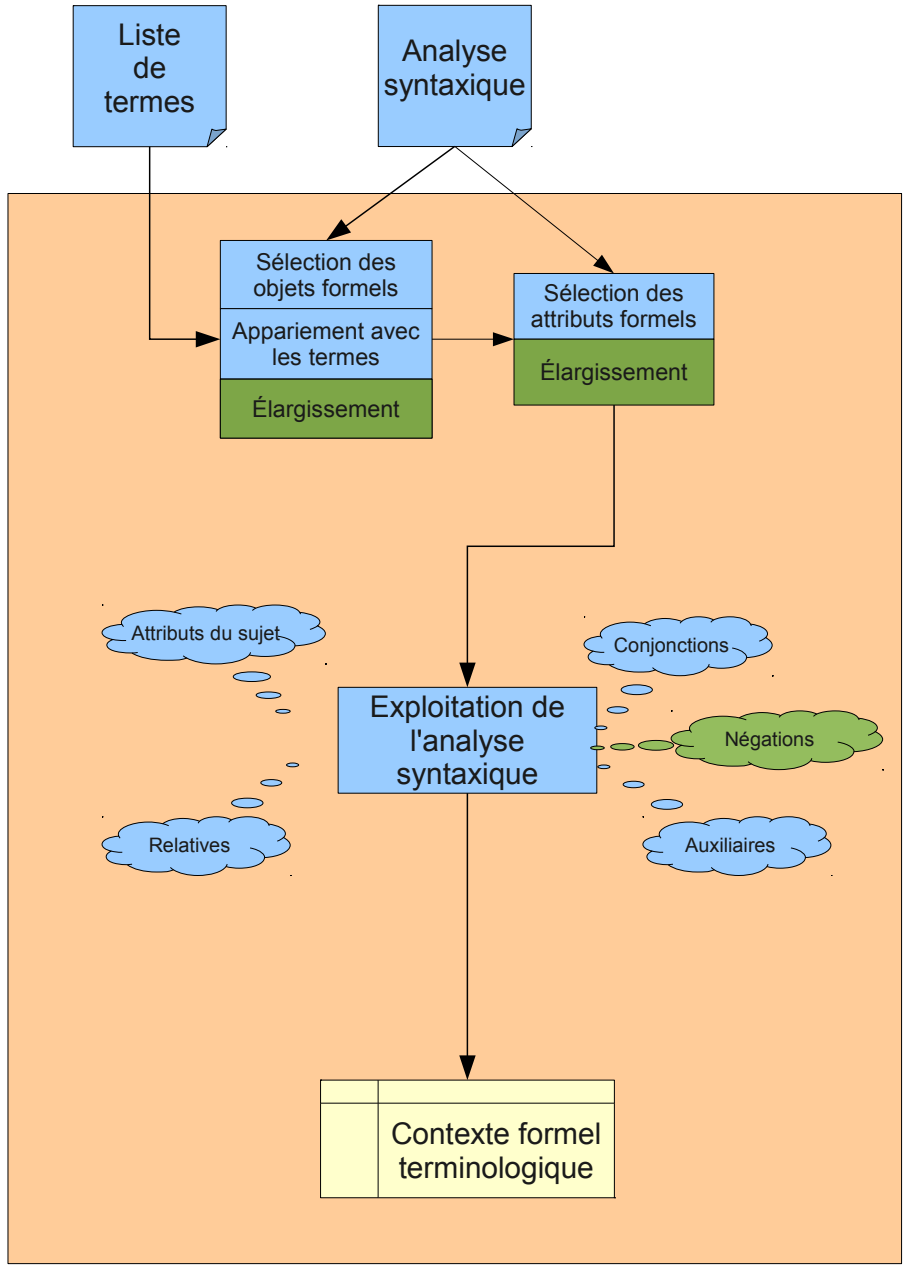
Produire un treillis interprétable : bilan

Améliorer l'interprétabilité du treillis selon deux axes

- Des éléments plus fidèles au texte
  - Attributs étendus
  - Objets étendus
  - Négations
- Un treillis mieux structuré
  - Ajout de connaissances
  - Séparation génériques / spécifiques

Méthodologie proposée





Sommaire

Pertinence des attributs de regroupement

Problématique

- L'ACF permet de regrouper des termes, mais pour de bonnes ou de mauvaises raisons ?
- Quelle est la proportion d'attributs formels pertinents ?

Attribut pertinent pour le domaine : évoque une propriété, une notion spécifique au domaine

Démarche

- La liste des 24 termes utilisés pour construire le corpus sur Wikipedia
- Un treillis construit à partir de cette liste (+ négation, attributs formels étendus)
- Une évaluation manuelle de la pertinence des attributs regroupant (dispersion$>$0)

Pertinence des éléments du treillis vis-à-vis du domaine (2)

Nombre d'objets	24
Nombre d'attributs	720
Dont négatifs	*54*
Nombre de concepts	104
Hauteur du treillis	6

Niveau	**Pert.**	**Tot.**	**Précision**
5 (+spécifique)	19	42	45,24%
4	10	20	50,00%
3	1	3	33,33%
2	1	2	50,00%
1 (+général)	0	3	0,00%
Moy.			**44,28%**

Pertinence des attributs de regroupement

- Plus on remonte dans le treillis, moins les attributs sont pertinents
- Ex. niveau 5 : « être étoile massive, être objet céleste »
- Ex. niveau 1 : « être en, posséder, former-ABLE »

Pertinence des éléments du treillis vis-à-vis du domaine (2)

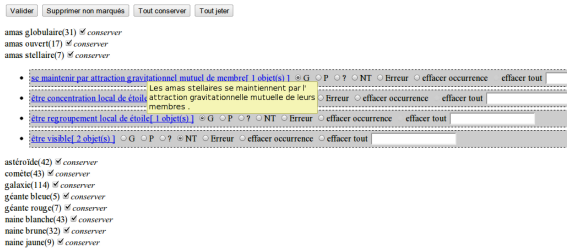
Nombre d'objets	24
Nombre d'attributs	720
Dont négatifs	*54*
Nombre de concepts	104
Hauteur du treillis	6

Niveau	Pert.	Tot.	Précision
5 (+spécifique)	**19**	**42**	**45,24%**
4	10	20	50,00%
3	1	3	33,33%
2	1	2	50,00%
1 (+général)	**0**	**3**	**0,00%**
Moy.			**44,28%**

Pertinence des attributs de regroupement

- Plus on remonte dans le treillis, moins les attributs sont pertinents
- **Ex. niveau 5 : « être étoile massive, être objet céleste »**
- **Ex. niveau 1 : « être en, posséder, former-ABLE »**

Taux de généricité des attributs propres



Problème : proportion d'attributs propres génériques et spécifiques ?

Exemples

- `se composer de trois parties` est générique pour `comète`
- `percuter Jupiter` est spécifique pour `comète`

49% des 273 couples (objet, attribut pertinent) sont génériques, **51%** sont spécifiques.

Bilan : il est essentiel de séparer les attributs génériques et spécifiques

Qualité des regroupements

Idée : confronter l'ACF à une ontologie de référence

Ontologie de référence

- Construite à partir du corpus d'astronomie de Wikipedia
- Tâche (fictive) : les objets célestes et leur devenir, classification et prédiction
- 73 concepts, quelques relations et instances

Des regroupements de référence

- Correspondent aux noeuds de l'ontologie et à leurs descendants qui apparaissent dans les 24 termes
- Sept regroupements de référence non disjoints

Procédure : Compter les recouvrements ensemblistes des regroupements de l'ACF avec ceux de référence

Qualité des regroupements (2)

Regroupements de référence

	Regroupement
R1	astéroïde, comète, naine brune, planète, satellite naturel
R2	amas stellaire, amas globulaire, amas ouvert
R3	galaxie, amas stellaire, amas globulaire, amas ouvert, quasar, étoile binaire
R4	étoile, géante rouge, géante bleue, naine jaune, naine rouge, supergéante rouge
R5	trou noir, pulsar
R6	supernova, nova
R7	naine blanche, naine noire

	Quantité	Dont pert.
$C = R$	0	0
$C \subset R$	10	9
$C \supset R$	4	0
C proche R	21	13
$C \neq R$	45	

- Pas de recouvrement exact
- Tendance à sur-généraliser mais à cause d'attributs non pertinents

Qualité des regroupements (2)

Regroupements de référence

	Regroupement
R1	astéroïde, comète, naine brune, planète, satellite naturel
R2	amas stellaire, amas globulaire, amas ouvert
R3	galaxie, amas stellaire, amas globulaire, amas ouvert, quasar, étoile binaire
R4	étoile, géante rouge, géante bleue, naine jaune, naine rouge, supergéante rouge
R5	trou noir, pulsar
R6	supernova, nova
R7	naine blanche, naine noire

	Quantité	Dont pert.
$C = R$	**0**	**0**
$C \subset R$	10	9
$C \supset R$	4	0
C proche R	21	13
$C \neq R$	45	

- **Pas de recouvrement exact**
- Tendance à sur-généraliser mais à cause d'attributs non pertinents

Qualité des regroupements (2)

Regroupements de référence

	Regroupement
R1	astéroïde, comète, naine brune, planète, satellite naturel
R2	amas stellaire, amas globulaire, amas ouvert
R3	galaxie, amas stellaire, amas globulaire, amas ouvert, quasar, étoile binaire
R4	étoile, géante rouge, géante bleue, naine jaune, naine rouge, supergéante rouge
R5	trou noir, pulsar
R6	supernova, nova
R7	naine blanche, naine noire

	Quantité	Dont pert.
$C = R$	0	0
$C \subset R$	10	9
$C \supset R$	**4**	**0**
C proche R	21	13
$C \neq R$	45	

- Pas de recouvrement exact
- **Tendance à sur-généraliser mais à cause d'attributs non pertinents**

ACF vs distributionnel

Objectif : positionner l'ACF par rapport à une méthode de classification hiérarchique

Méthodologie

- Une méthode d'analyse distributionnelle sur des termes sans syntaxe [Bannour et al., 2011]
- Classification hiérarchique ascendante (résultat le plus favorable)
- Référence : les 24 termes initiaux étendus (trou noir supermassif...)
- Liste de termes pour l'ACF : tous les termes des regroupements de référence.

	ACF	Dis.	ACFb
$C = R$	0	2	1
$C \subset R$	8	3	10
$C \supset R$	0	0	1
C proche R	11	1	7
$C \neq R$	50	3	50

- ACF plus bruité
- ACF plus de regroupements partiels
- Les subsomptions explicites améliorent les résultats

ACF vs distributionnel

Objectif : positionner l'ACF par rapport à une méthode de classification hiérarchique

Méthodologie

- Une méthode d'analyse distributionnelle sur des termes sans syntaxe [Bannour et al., 2011]
- Classification hiérarchique ascendante (résultat le plus favorable)
- Référence : les 24 termes initiaux étendus (trou noir supermassif...)
- Liste de termes pour l'ACF : tous les termes des regroupements de référence.

	ACF	Dis.	ACFb
$C = R$	0	2	1
$C \subset R$	8	3	10
$C \supset R$	0	0	1
C proche R	11	1	7
$C \neq R$	**50**	**3**	50

- **ACF plus bruité**
- ACF plus de regroupements partiels
- Les subsomptions explicites améliorent les résultats

ACF vs distributionnel

Objectif : positionner l'ACF par rapport à une méthode de classification hiérarchique

Méthodologie

- Une méthode d'analyse distributionnelle sur des termes sans syntaxe [Bannour et al., 2011]
- Classification hiérarchique ascendante (résultat le plus favorable)
- Référence : les 24 termes initiaux étendus (trou noir supermassif...)
- Liste de termes pour l'ACF : tous les termes des regroupements de référence.

	ACF	Dis.	ACFb
$C = R$	0	2	1
$C \subset R$	**8**	**3**	10
$C \supset R$	0	0	1
C proche R	11	1	7
$C \neq R$	50	3	50

- ACF plus bruité
- **ACF plus de regroupements partiels**
- Les subsomptions explicites améliorent les résultats

ACF vs distributionnel

Objectif : positionner l'ACF par rapport à une méthode de classification hiérarchique

Méthodologie

- Une méthode d'analyse distributionnelle sur des termes sans syntaxe [Bannour et al., 2011]
- Classification hiérarchique ascendante (résultat le plus favorable)
- Référence : les 24 termes initiaux étendus (trou noir supermassif...)
- Liste de termes pour l'ACF : tous les termes des regroupements de référence.

	ACF	Dis.	**ACFb**
$C = R$	0	2	**1**
$C \subset R$	8	3	**10**
$C \supset R$	0	0	**1**
C proche R	11	1	**7**
$C \neq R$	50	3	**50**

- ACF plus bruité
- ACF plus de regroupements partiels
- **Les subsomptions explicites améliorent les résultats**

Sommaire

Conclusion

Mes apports

- Intégration des termes dans l'ACF
- Analyse fine du comportement de l'ACF sur les textes
- Analyse du comportement des termes et des entités nommées
- Paramétrage de l'ACF pour améliorer l'interprétabilité du treillis

Apports de l'ACF pour la COT

- Le treillis doit être considéré avec prudence
  - L'ACF ne produit pas une ontologie
  - Les regroupements sont très bruités : inutilisables tels quels
- **Le treillis peut être utilisé comme une représentation synthétique du texte**
- **Le treillis peut être interprété en ontologie**
  - Il faut séparer les attributs génériques/spécifiques pour obtenir des regroupements cohérents
  - C'est un processus manuel coûteux

Perspectives

- Augmenter encore l'interprétabilité du treillis
  - Utiliser plus finement l'analyse syntaxique
  - Injecter plus de connaissances
- Faciliter la séparation des attributs génériques/spécifiques
  - Repérage de marqueurs dans le texte
  - Apprentissage supervisé pour amorcer le processus
- Mener une évaluation orientée utilisateur
- Étudier des stratégies de visualisation/navigation dans le treillis

Bannour, S., Audibert, L., & Nazarenko, A. (2011).
Mesures de similarité distributionnelle entre termes.
In *Actes des 22èmes Journées francophones d'Ingénierie des Connaissances (IC2011)* Chambéry, France.

Bendaoud, R., Toussaint, Y., & Napoli, A. (2008).
Pactole : A methodology and a system for semi-automatically enriching an ontology from a collection of texts.
In P. W. Eklund & O. Haemmerlé (Eds.), *Proceedings of the 16th International Conference on Conceptual Structures (ICCS 2008)*, volume 5113 of *Lecture Notes in Computer Science* (pp. 203–216). : Springer.

Cimiano, P., Hotho, A., & Staab, S. (2005).
Learning concept hierarchies from text corpora using formal concept analysis.
Journal of Artificial Intelligence Research (JAIR), 24, 305–339.

Cimiano, P. & Völker, J. (2005).
Text2Onto — a framework for ontology learning and data-driven change discovery.
In A. Montoyo, R. Munoz, & E. Metais (Eds.), *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 3513 of *Lecture Notes in Computer Science* (pp. 227–238). Alicante, Spain : Springer.

Fortuna, B., Grobelnik, M., & Mladenic, D. (2006).
Semi-automatic data driven ontology construction system.
In *Proceedings of the 9th International multi-conference Information Society IS-2006, Ljubljana, Slovenia*.

Ganter, B. & Wille, R. (1999).
Formal Concept Analysis.
Berlin : Springer.

Harris, Z., Gottfried, M., Ryckman, T., Mattick, P. J., Daladier, A., Harris, T., & Harris, S. (1989).
The Form of Information in Science, Analysis of Immunology Sublanguage, volume 104 of *Boston Studies in the Philosophy of Science*.
Boston : Kluwer Academic Publisher.

Hearst, M. A. (1992).
Automatic acquisition of hyponyms from large text corpora.
In *Proceedings of the 14th conference on Computational linguistics-Volume 2* (pp. 539–545). : Association for Computational Linguistics.