Soutenance de HDR

Intégration structurale des points de vue componentiels et compositionnels : pourquoi et comment

I] Parcours
II] Exposé

Dominique DUTOIT
Memodata, Crisco, Litis

Parcours

Professionnel

Gérant de Memodata depuis 1989

Construction de dictionnaires

Étude des applications des dictionnaires

D'enseignement

4000 heures de face à face pédagogique

Informatique – TAL – SI – Méthodologie – Gestion

De recherche

14 contrats de recherche

21 publications

(+9 acceptées non publiées)

Intégration structurale
des points de vue
componentiels et compositionnels
pourquoi et comment

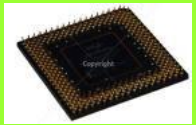
Dominique DUTOIT
Memodata, Crisco, Litis

Une problématique de Signes linguistiques

Sens des Significations

Signification des textes
(commentaire à propos des signes)

Sens des Significations



Problématique de la Sémantique Générale : *hors sujet.*

Problématique extensionnelle linguistique de la Thèse (TST, théorie componentielle etc.)

Problématique intensionnelle linguistique de la HDR.

Extensionnalité/intensionnalité

1°)	**Le rayon *condiment* de mon supermarché**	**Ling.**	**Nécessaire extensionalité**
2°)	La théorie formelle XXX a été un condiment intéressant de la linguistique du 20^{e} s.	Ling.	Nécessaire intensionnalité
3°)	Quel est le condiment principal du lapin à la moutarde?	Ling.	Nécessaire articulation

Terme de cuisine

ayant un « *sens* »

quel en est le goût?

Le dictionnaire et le corpus

dictionnaire : ouvrage à usage du lecteur ***humain***.
tentant une description ***universaliste*** et ***apriorique*** du lexique tel qu'il apparaît dans les **corpus**

Pourquoi le dictionnaire rend-il service à l'humain?
du fait de la "créativité" de ce dernier.

"Créativité" due à quoi?
aux processus interprétatifs eux-mêmes
Lesquels sont assez inconnus…

D'où deux tâches comme **deux mâchoires** :

Formalisation du dictionnaire	*Par et pour les textes…*
Par et pour la formalisation de processus « créatifs » propres aux éléments du lexique.	**Mise à l'épreuve au cours de l'analyse automatique des textes**

Des opérations linguistiques

Opérations traitées en extension

- Opération texte → sens

Désambiguïsation lexico-sémantique

Semio 1

- Plusieurs Opérations texte→sens→texte

Dictionnaire à l'envers (réduction lexicale) - Résumé lexical, thématique - paraphrases d'énoncés courts - filtrage d'information, ajout de co-texte, signature sémantique ...

Semio 1

Opérations traitées en intension

- Opérations intensionnelles texte → sens

Désambiguïsation lexico-sémantique

Semio 2

- Opérations intensionnelles texte→sens→texte

QR, coréférence et extraction d'information.

Semio 2

La démarche de présentation

Sémiographe V1 (chapitre 1 à 5)

- Les données : le Dictionnaire Intégral
- Les traitements sémantiques : le Sémiographe
- Une Application : production d'Alexandria

Sémiographe V2 (chapitre 6)

- La structure : Dictionnaire & corpus
- L'attention
- Le temps

Le Dictionnaire Intégral

- Description qualitative
  - TST, Sémantique componentielle (LDI), WordNet, ~~SUMO~~ , ~~CYC,~~ *Framenet*
  - Complémentarité des deux théories, framenet, autres dictionnaires
- Description quantitative et applicative
  - Alexandria

TST

TST : produire **toutes** les paraphrases d'un "sens".

La TST repose sur un dictionnaire organisé selon des fonctions lexico-sémantiques :

S_0(éclipser) = *éclipse* (nominalisation), S_1(éclipser) = *corps céleste*.

Ces fonctions n'expliquent pas les intensions associées aux lexies.

Par exemple, un même sens pourra produire :

1°) La lune éclipse le soleil et *2°) l'éclipse du soleil par la lune*

Mais non :

3°) *la disparition du Soleil causée par l'apparition d'un grand disque le cachant…*

qui nécessite précisément une prise en compte de l'intension

conclusion:

le but **TOUTES** ne peut être atteint par cette théorie.

SEMantique COmponentielle

La sémantique componentielle décompose les significations en traits plus élémentaires (sèmes) et espère, depuis cette décomposition, découvrir des lois de composition.

TraitSém(éclipse) = [abstrait] [cacher] [action] [fait]…

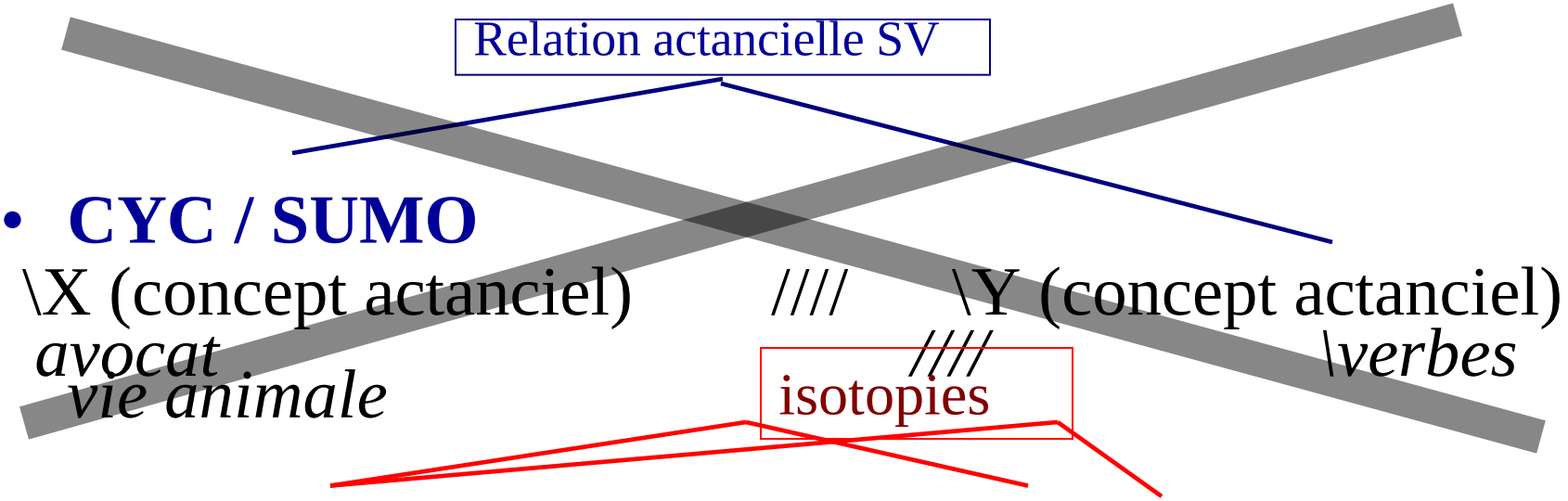
Par exemple, elle appariera assez bien:

2°) l'éclipse du soleil par la lune **et** 3°) *la disparition du Soleil causée par l'apparition d'un grand disque le cachant…*

Mais produira quelques bruits…

LDI

Les relations



- **CYC / SUMO**

\X (concept actanciel) //// \Y (concept actanciel)

avocat //// *verbes*

vie animale

- **SemCo**

\X (trait de sens, concept) //// \Y (concept)

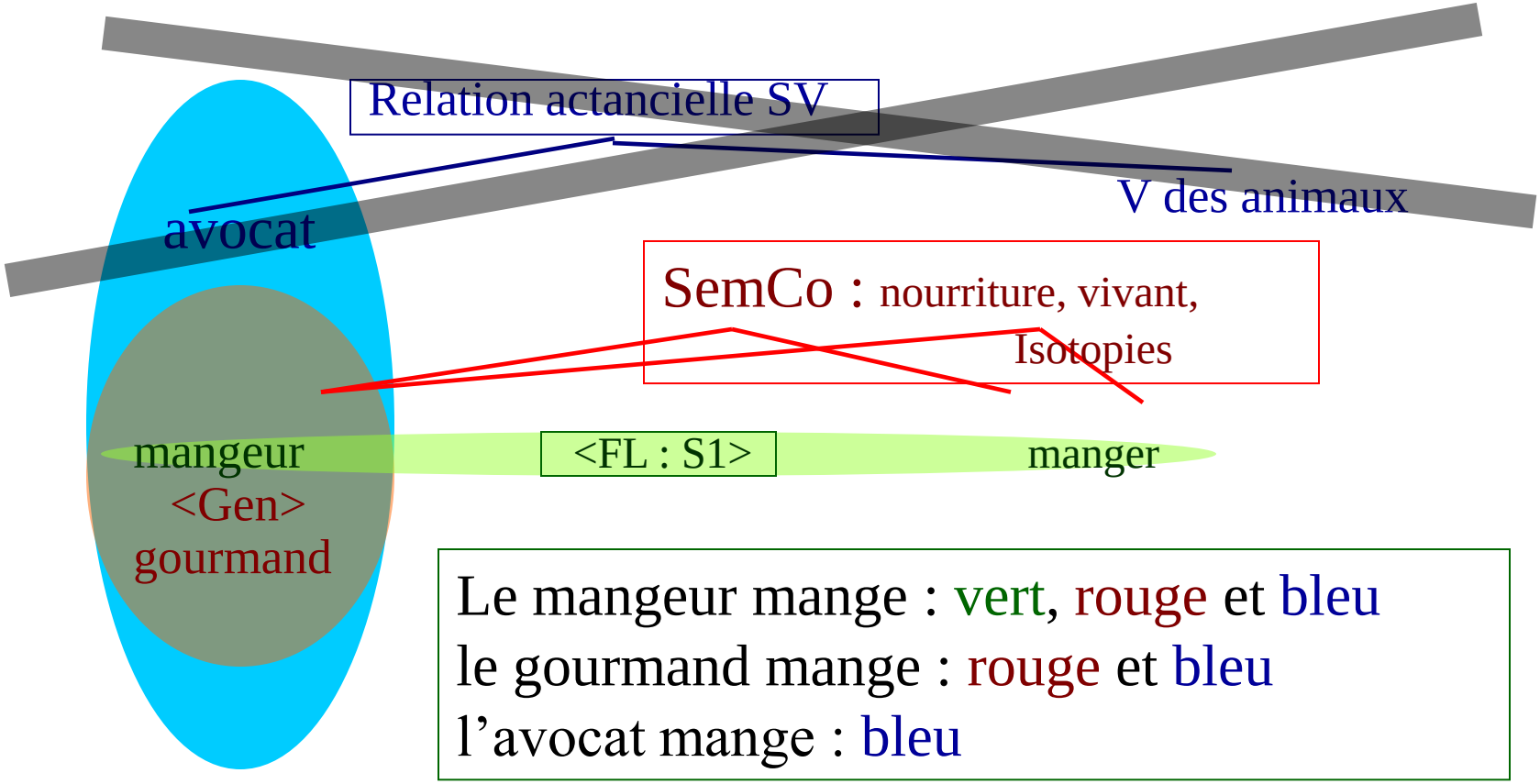
gourmand=Spec mangeur //// *\nourriture, manger*

- **TST**

X (mot-sens ou sens) <FL> Y (mot-sens ou sens)

mangeur <FL> manger

Graphique complet



Le mangeur mange : vert, rouge et bleu
le gourmand mange : rouge et bleu
l'avocat mange : bleu

Complémentarité des deux théories, framenet, autres dict.

$f_{TST}(couleur, surface) = \{\}$ $f_{SemCo}(couleur, surface) = \{[surface]\}$

$f_{TST}(colorer, rouge) = \{S_{mod}\}$ $f_{SemCo}(colorer, rouge) = \{[couleur]\}$

Complémentarité	Limite
TST → production de paraphrases SEMCO → analyse	on ne trouve pas de lieux partagés entre les deux espaces.
Framenet (grammaire de cas)	
$f_{Framenet}(couleur, X) = \{[X=entity]\}$ *... Pas très précis, mais déjà qqch.*	
Le Robert (dictionnaire ordinaire)	
$f_{Robert}(couleur, surface) = \{def(couleur)=$ *caractère de la surface d'un objet*$)\}$.	Précis, mais quelle est le mode d'emploi
Cambridge (dictionnaire ordinaire)	
$f_{cambridge}(couleur, surface) = \{def(color)=$ *appearance of sth reflecting light*$)\}$.	

Architecture

Selon le précepte cartésien, pour réduire la complexité de l'analyse, on morcelle le problème en un ensemble de problèmes réputés plus simples (Enjalbert, 2005, page 271) :

- Tokenisation
- (expressions à figement) (1)
- **Étiquetage**
- (Analyse syntaxique) (2)
- **Analyse lexico-sémantique**

En gras : présentation ci-après

1 et 2 : modules généralement désactivés.

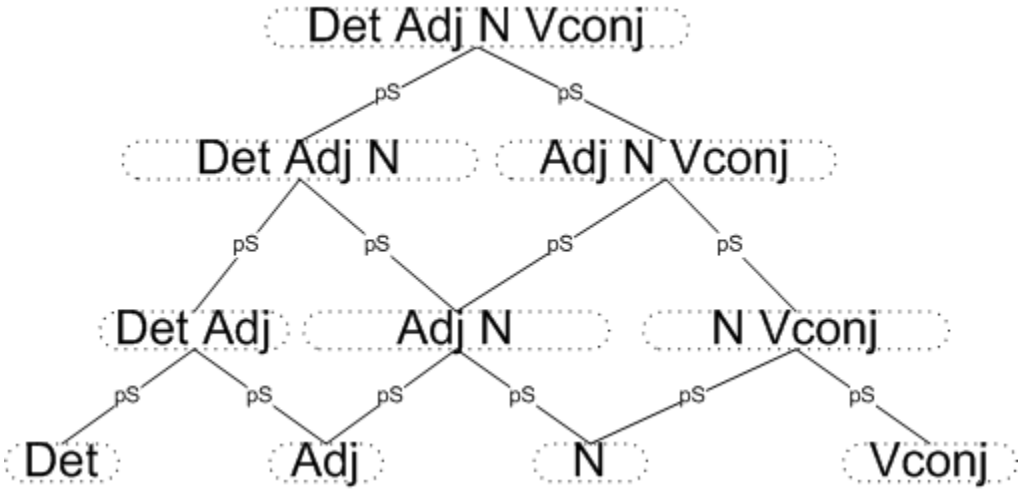
Tout l'enjeu du Semio version 2 et du chapitre 6 de la HDR est de faire disparaître ces niveaux d'analyse.

Etiquetage

S'il existe

Cet obséquieux personnage avait…

alors les séquences suivantes existent nécessairement en français:



Technique statistique depuis corpus non étiqueté.

WSD

Les “isotopies” (éléments partageant les mêmes traits sémantiques)

→ pas de connaissances courantes ni de TST

→ pas de différence de traitement syntagme/espace hors syntagme

→ des énoncés non solutionnables comme

l’avocat mange un avocat

Une application

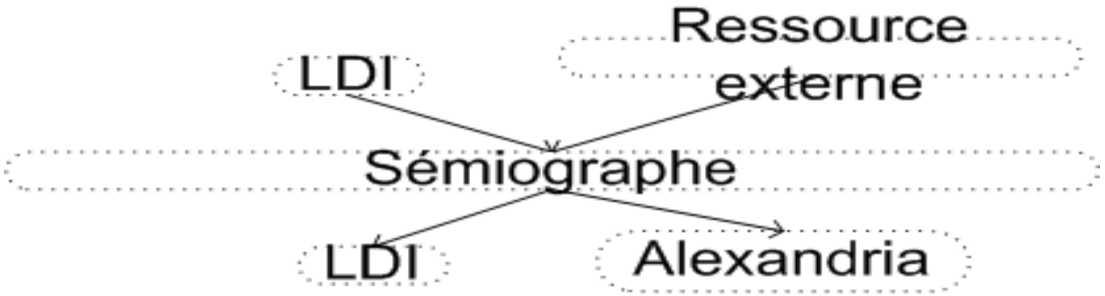
Aide à la construction d'Alexandria.

Alexandria :

- BDD lexicale. 26 langues,
- 2 millions de relations

Usage du semiographe pour Alexandria:

- Fusion (aide à la) définitions fr. avec les mots-sens fr de LDI
- Fusion synset en wordnet / synset fr LDI.



Conclusion

Algorithme et ressources utilisées

Base de données exploitée

- Alexandria : 100.000 visiteurs par jour
- Nombreuses intégrations industrielles

Mais,

Développement incomplet : intension.

Sémiographe v2

1°) Nous avons noté qu'au niveau du syntagme
notaire mange est

- **Inaccessible** depuis TST, SemCo, FrameNet, Dict…
- a fortiori, non modélisé en terme **informationnel**.

2°) La chapitre 6 de la HDR montre pourquoi il y a obligation de rompre avec l'isolement de ces deux questions.
A l'aide d'exemples élémentaires, partant de la forme stricte et supposée sans *sens* et terminant par l'étude d'une question simple, il défend la nécessité de maintenir les quatre unités suivantes:

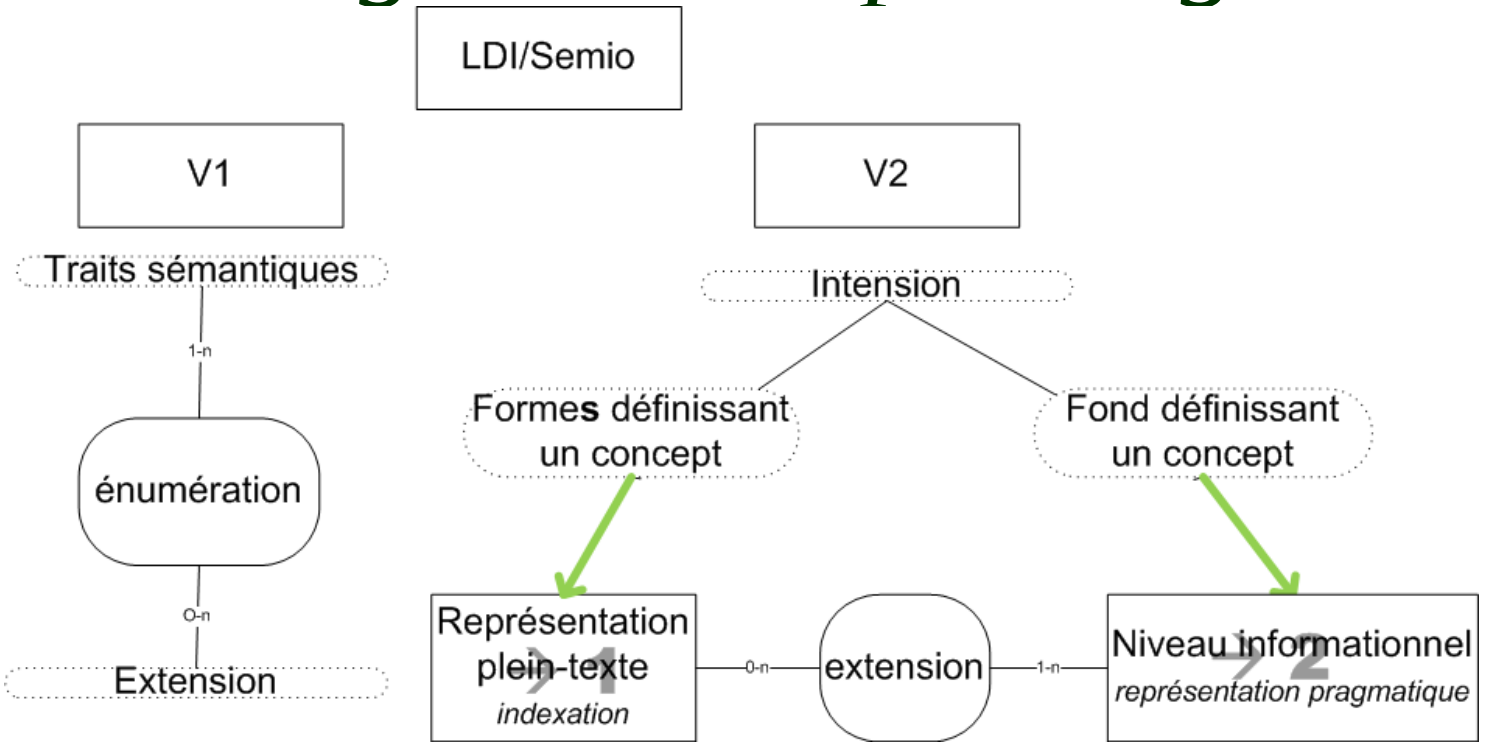
de Structure – d'Analyse – de Résultat – du Signe

Plan

Définition - Problématique / Indexation, structure méréologique

Exemples , applications / Génération et Méta –représentation / Méréotopologie, Grammaire fonctionnelle, intégrations.
Conclusion

Changement de paradigme



théorie des ensembles méréo(topo)logie

Problématique

Considérant par exemple le *dictionnaire à l'envers*, l'indexation des définitions améliorerait la qualité des réponses en fournissant des dépendances ne serait-ce qu'au niveau des *séquences*.

En considérant Séquences, SemCo et TST,

En excluant module génératif et représentation intensionnelle de fond,

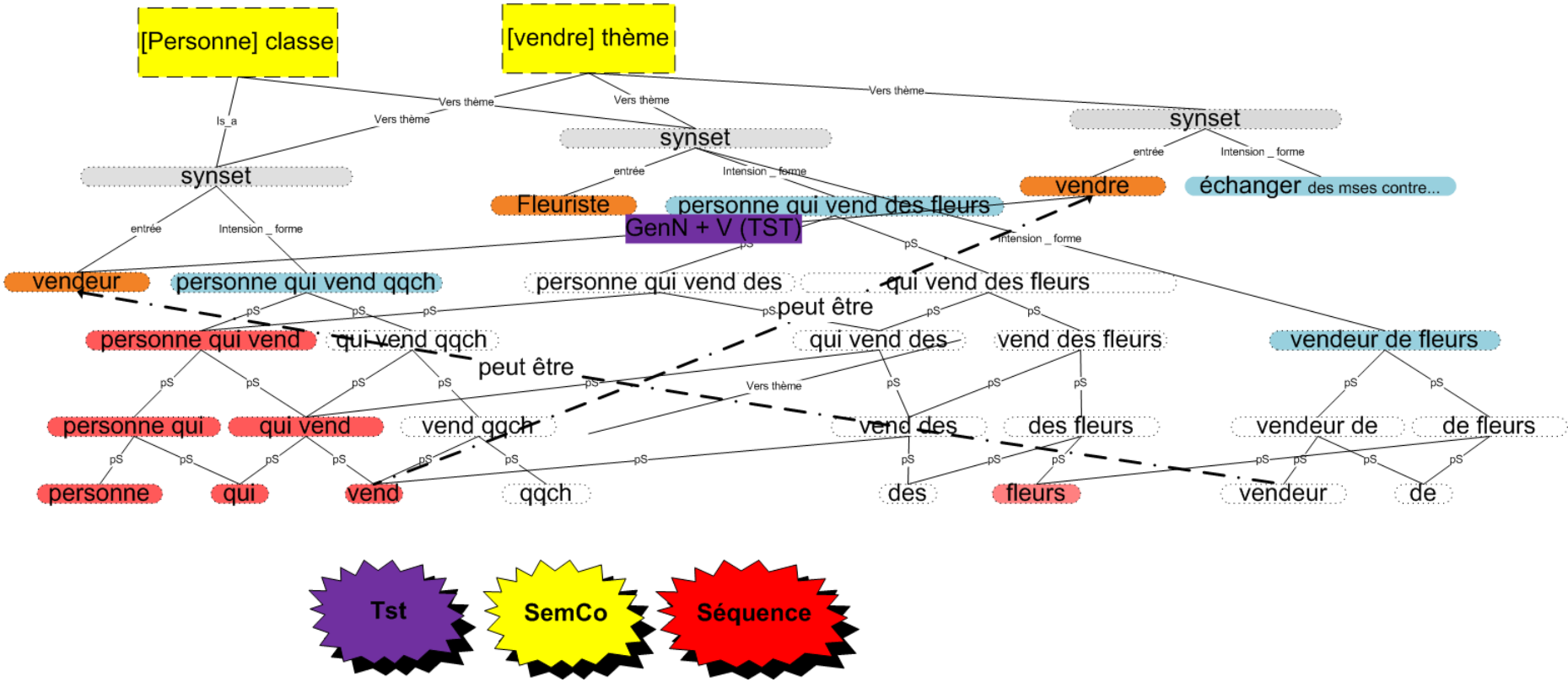
Représenter les lieux où les *définitions* suivantes se recoupent ?

Personne qui vend qqch

Personne qui vend des fleurs

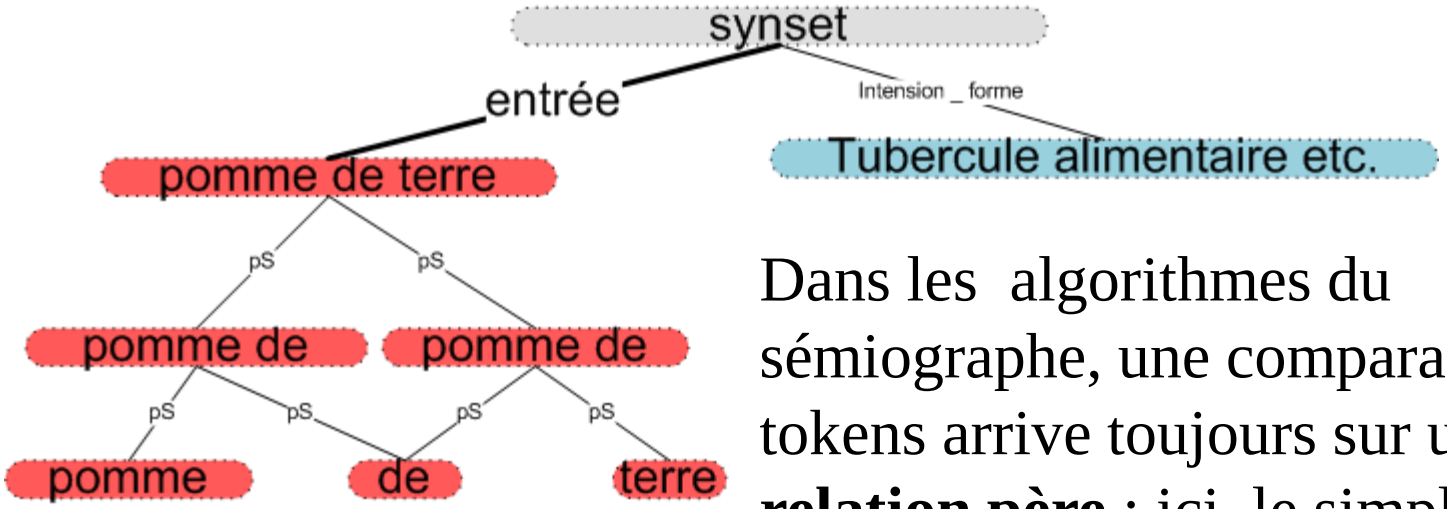
Vendeur de fleurs.

SemCo, TST, Séquenceur



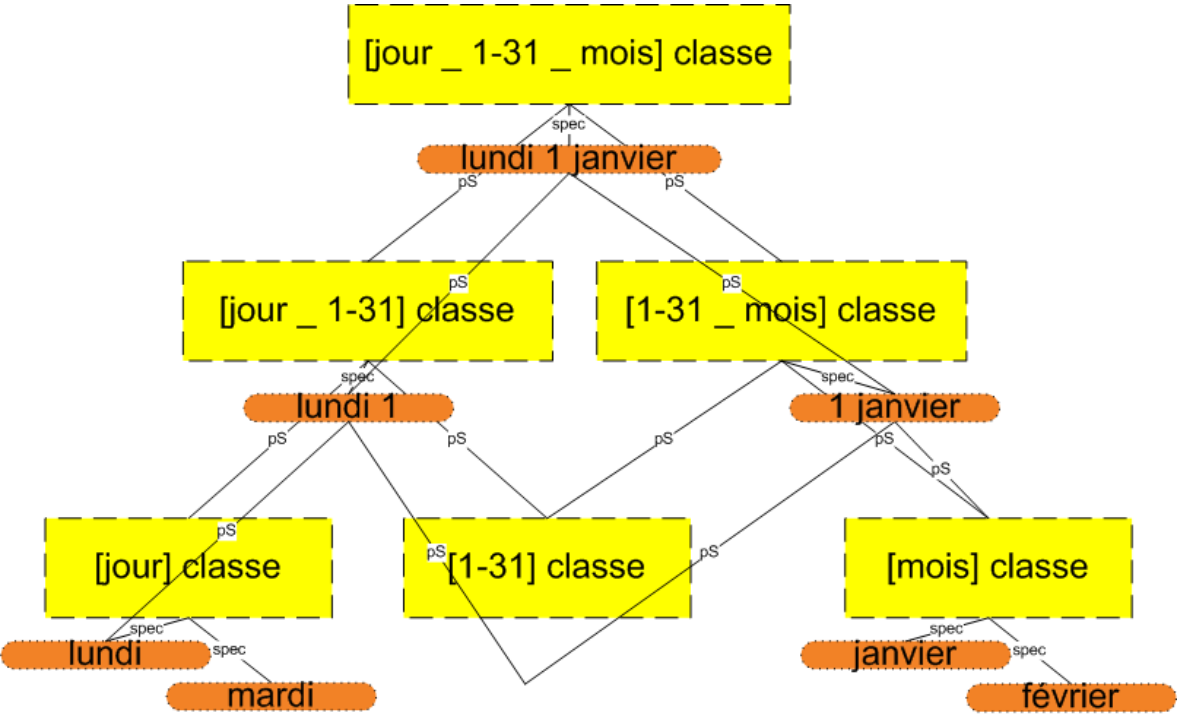
Quelques conséquences fonctionnelles

Recherche d'expressions plus ou moins figées (1/2)



Dans les algorithmes du sémiographe, une comparaison de n tokens arrive toujours sur une **relation père** : ici, le simple test de la relation « entrée » aboutit à **l'intégration légère d'un moteur de recherche des expressions.**

Recherche d'expressions plus ou moins figées (2/2)

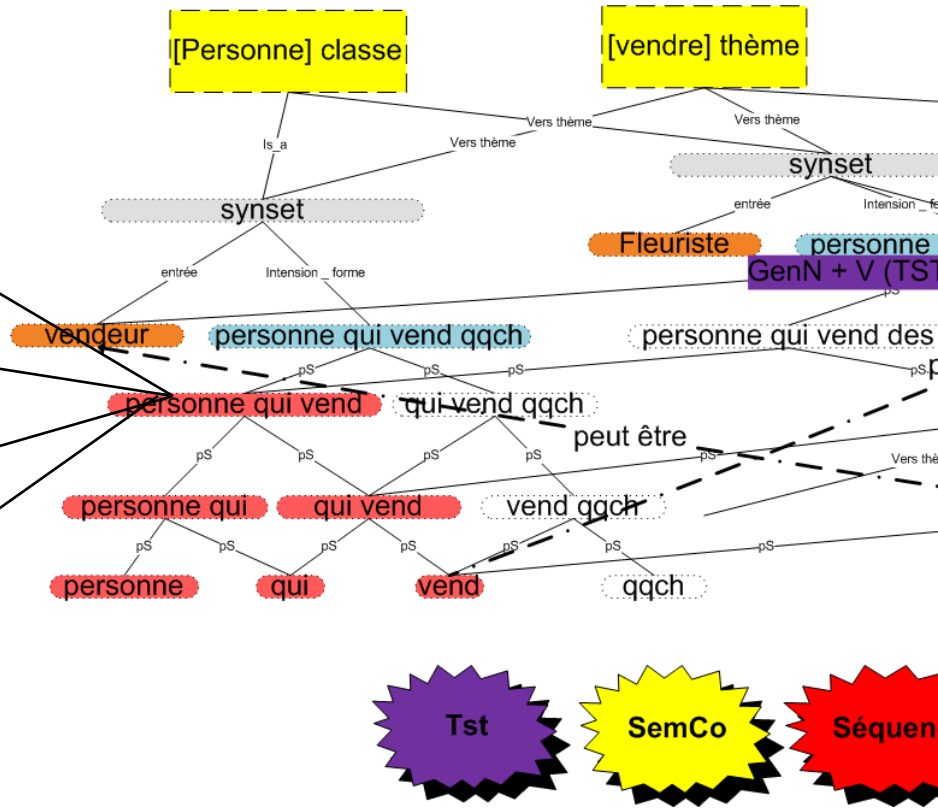


De toute façon, les séquences lundi 1, lundi 1 janvier etc. auraient été créées... Seule ajout, un résultat *spec*, tout à fait équivalent du résultat *entrée* précédent.

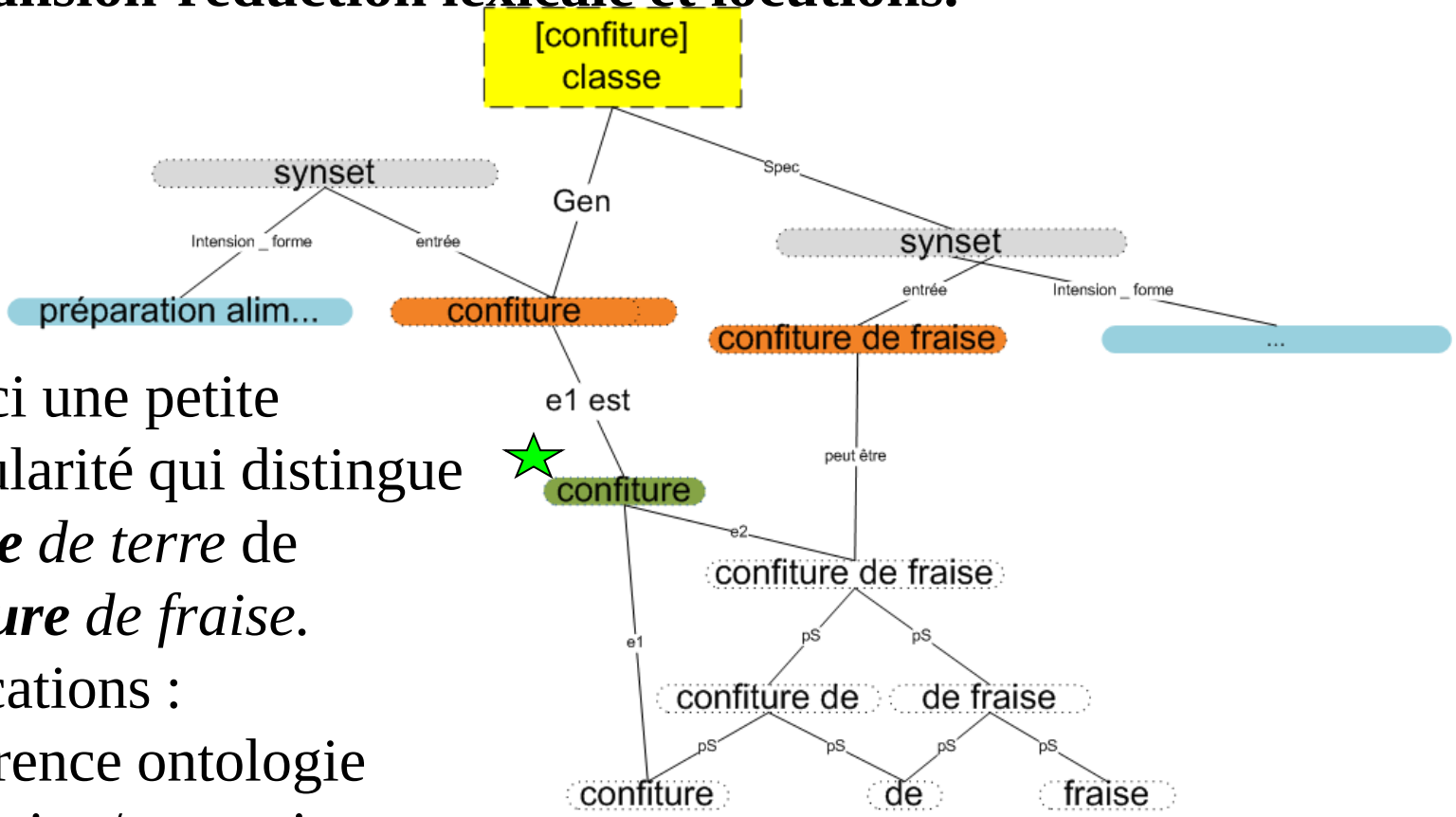
Ici, la date est à la fois retrouvée, normalisée, et réutilisable pour d'autres traitements.

Implicitement, un moteur d'indexation à la Google…

- je recherche une *personne qui vend* des patchwork qu'elle réalise **...**
- un rideau à réaliser avec des tissus que je fournirai - Autres - Rennes.
  *rennes.olx.fr/je-recherche-une-**personne-qui-vend**-des-patchwork-qu-elle-realise-iid-23105279* - En cache - Pages similaires -
- Cherche *Personne Qui Vend* Gta Sa Sur Pc (Plateforme PC) (Jeux **...**
- 2 messages - 2 auteurs - Dernier message : 16 mai 2008
- Cherche *Personne Qui Vend* Gta Sa Sur Pc (Plateforme PC) : Cherche Gta Sa sur Pc et personne non arnakeur avec jeux ki marche et ki habite en **...**
  *www.jeuxvideo.fr/forum/plateforme-pc/cherche-**personne-qui-vend**-gta-sa-sur-pc-plateforme-pc-id341212-page1.html* - En cache - Pages similaires -
- Recherche *personne qui vend* r4 pour nds : Forum auFeminin
- Recherche *personne qui vend* r4 pour nds. bonjour, je recherche un R4 nds pour mon fils et ma fille!!! merci valerie-go34@hotmail.fr -Vends carte r4 + jeux + **...**
  *forum.aufeminin.com/forum/f1224/__f12_f1224-Recherche-**personne-qui-vend**-r4-pour-nds.html* - En cache - Pages similaires -
- Cherche *personne qui vend* abonnement movida toulouse!!!! : Forum **...**



Identification (1/2). Expansion-réduction lexicale et locutions.



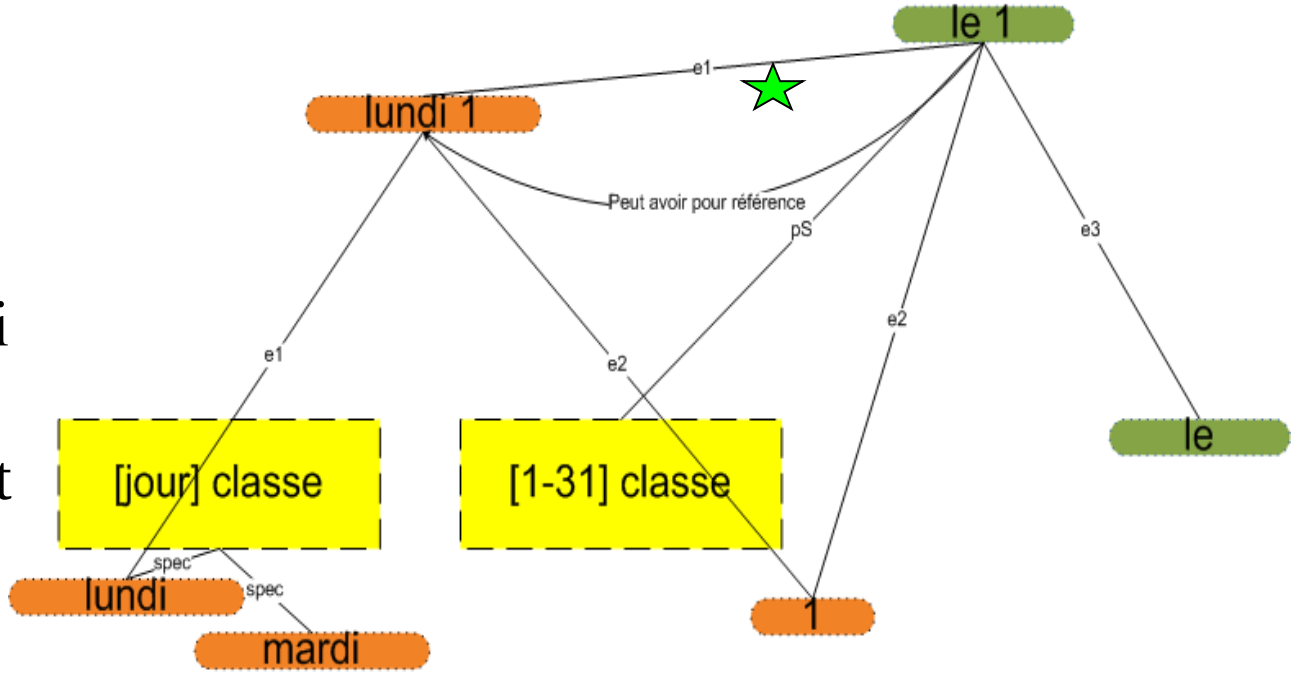
★Voici une petite particularité qui distingue ***pomme*** *de terre* de ***confiture*** *de fraise*.

Applications :

- cohérence ontologie
- réduction/expansion lexicale : traduction.

Identification (2/2). Référence – accès à des ellipses.

★Il y a ici l'emploi d'une **approche générative** pointant sur les ellipses. D'autres solutions sont-elles envisageables?



Conclusion

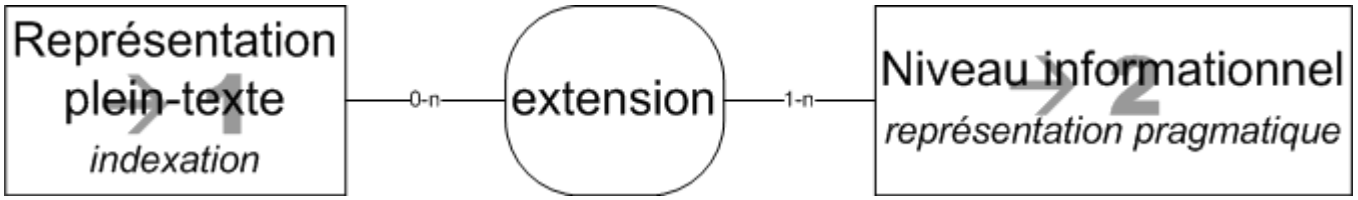
En élevant le statut de la simple séquence (en respectant une relation d'ordre partiel) au rang de « concept », nous avons **intégré d'une façon homogène***

- la détection des **locutions figées**
- la détection des **locutions à énumération**
- l'accès à une forme qualifiée par une **dénomination** en langue (« date » n'a nul besoin de métalangage)
- recréé un moteur d'indexation et de recherche proche de **Google**.

En complétant ces résultats au moyen de réécritures d'identification, nous avons :

- rendu accessible conditionnellement certaines **ellipses** (référence)
- augmenter la cohérence ***logique*** du graphe.

*MAIS 2 reste à ce point seulement une **forme membre d'une classe**.*
*Quel est **l'intension de fond** de cette classe?*



Notes préliminaires

1°) Cette partie fera l'économie de certains graphes dont on admettra qu'on ne peut pas plus les envisager, les concevoir dans leur entièreté qu'on ne peut concevoir dans son entièreté un cube (cf. *Bertrand Russel*).
2°) La relation d'arité 2 permettant la construction des séquences et le maintien aisé de la relation d'ordre est supposée toujours utilisée, quelque soit le niveau des représentations, même si, dans un souci d'allègement nous utilisons des relations d'arités supérieures.
Ainsi, à une arité de 3 correspond toujours 2, voire 3 groupes d'arité de 2, selon qu'un certain ordre compte ou non.

Problématique

En considérant Séquences, SemCo et TST, et un **module génératif** *les deux énoncés suivants sont très synonymes:*

Personne qui vend des fleurs

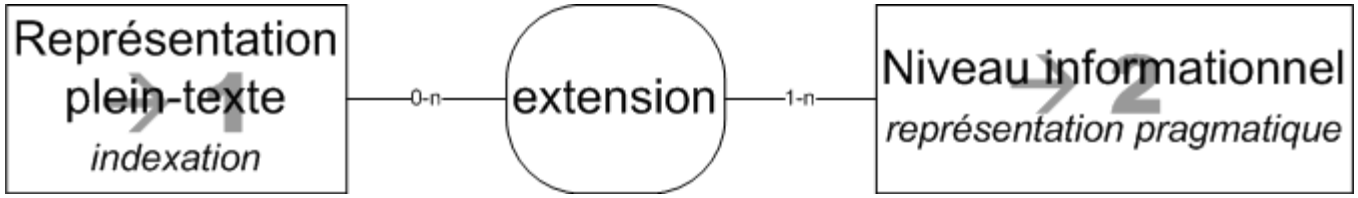
Vendeur de fleurs

Mais que valent chacun des mots de ces énoncés? Et que vaut l'ensemble considéré?

La *fleur* est-elle	nécessairement	la partie sexuée d'une plante?
Le *vendeur* est-il		une personne?
La *personne* est-elle		un humain?
Vendre est-ce		contre de l'argent?
Le *touts* sont-ils		un fleuriste?

Et si nous refusions plutôt toute **surinterprétation**?

Principaux concepts



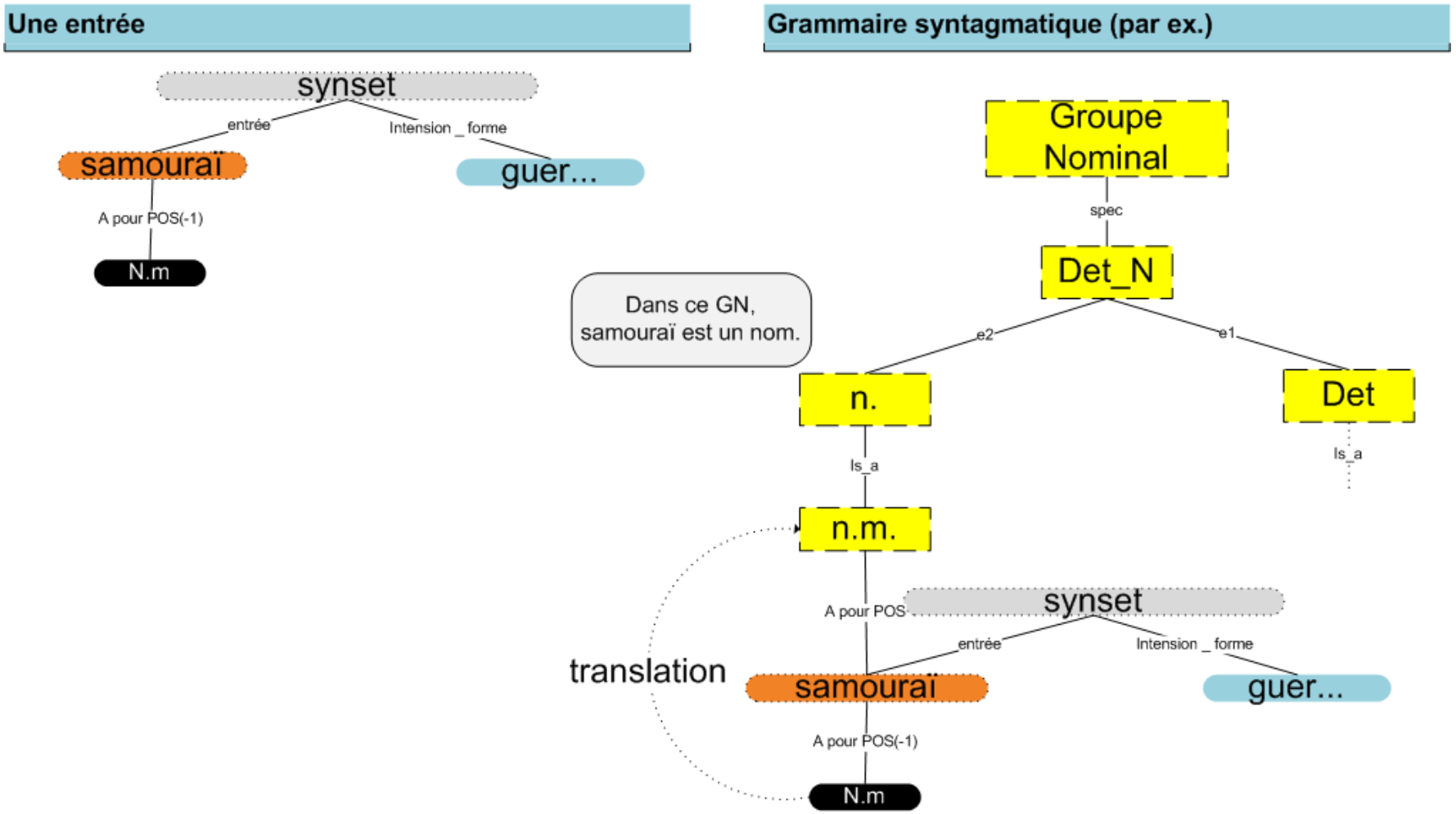
- puissance active (*essai*) / puissance passive (*monde*)
- définition

Exploration :

1°) le nom *samouraï* (existe), le nom *abattre* (n'existe pas en français).

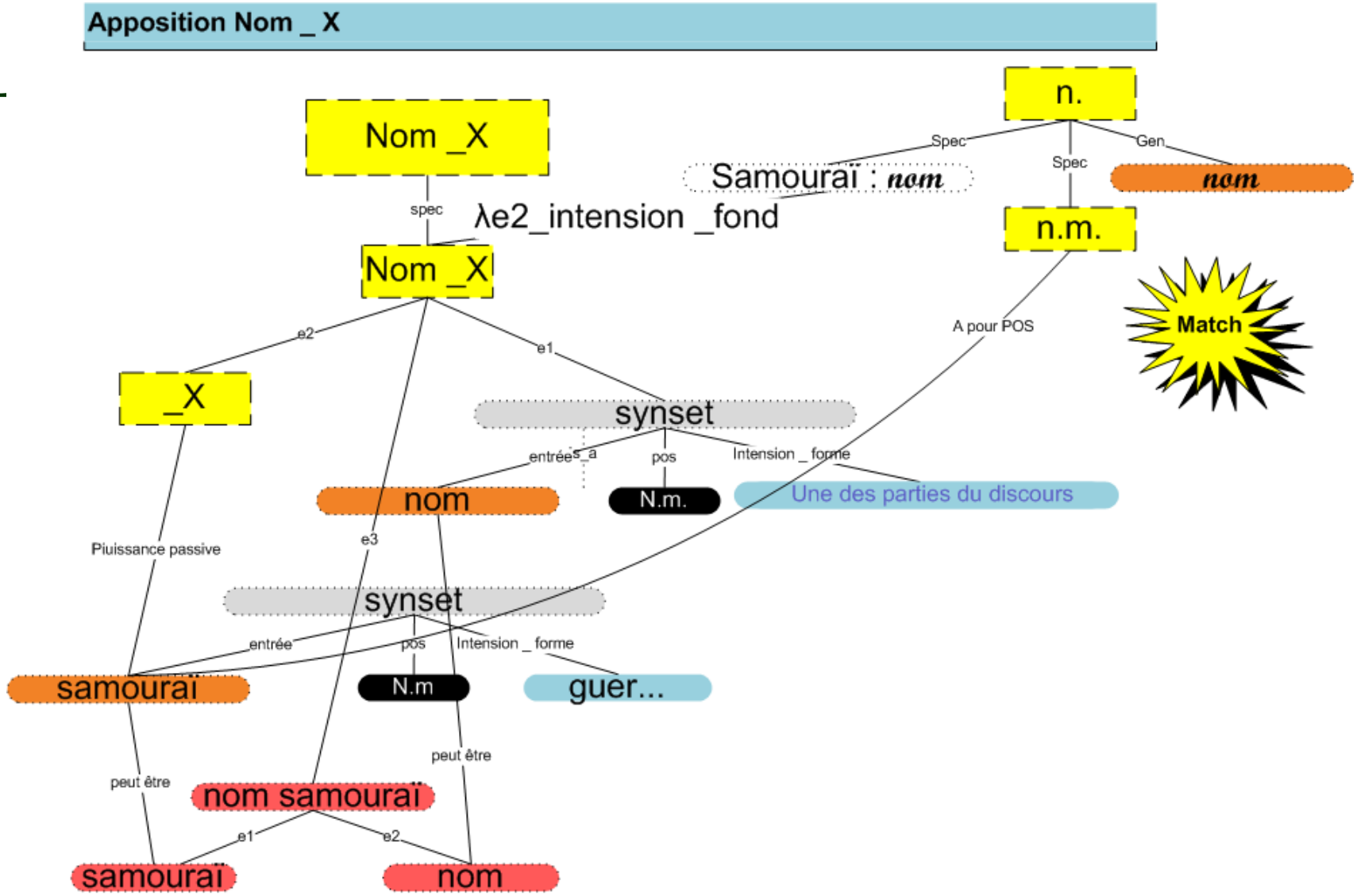
2°) Une histoire de *cheval blanc*

Le nom *samouraï*, le nom *abattre* (1/3).

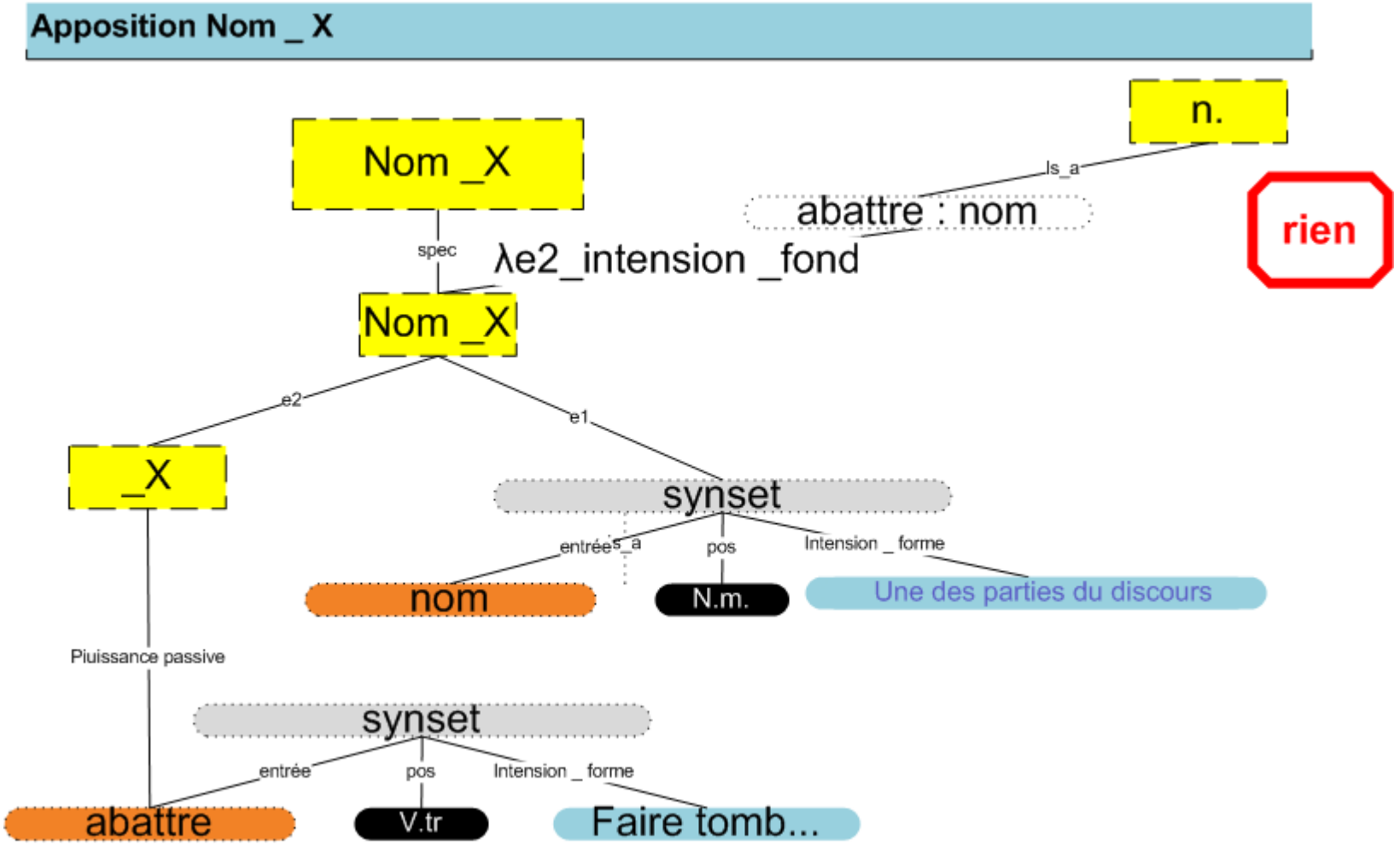


L'intension comme **fond**

L



Le nom *samouraï*, le nom *abattre* (2/3).

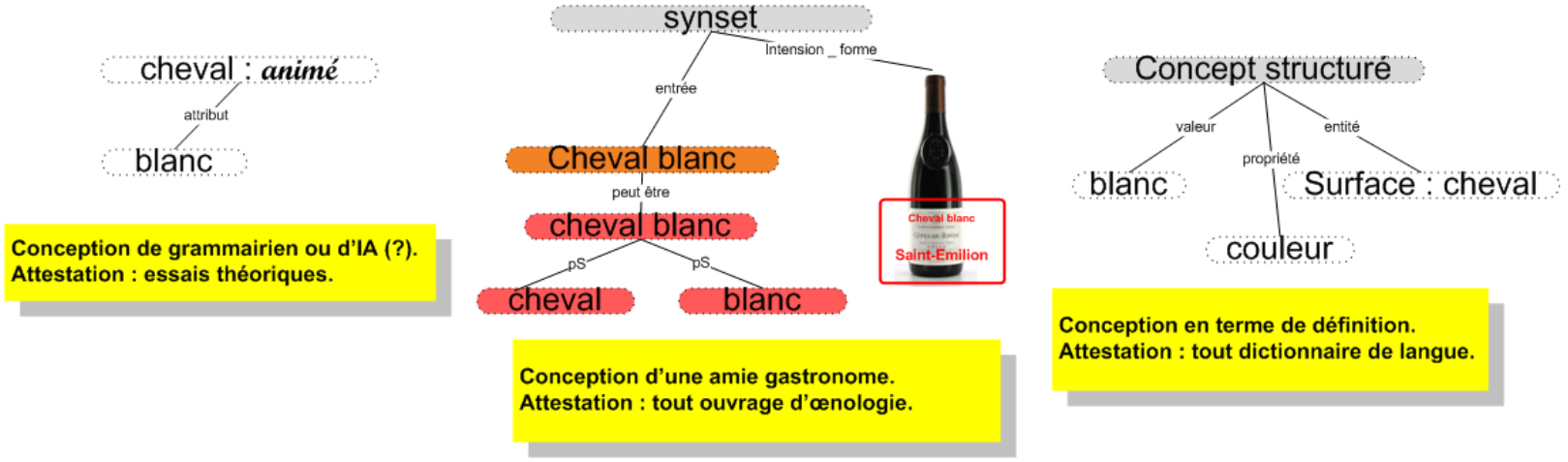


Définitions

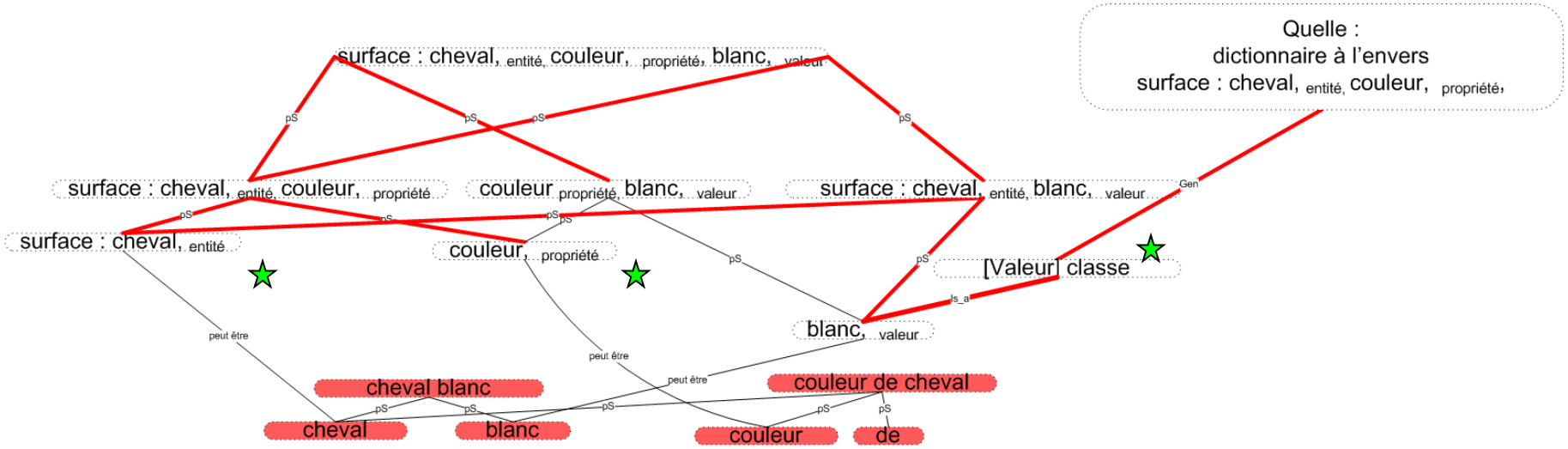
Puissance active : principe du mouvement, **tente** de déplacer ce qui est en dans sa sphère d'influence (fourni ici par la syntaxe), manifestation de qualités intrinsèques.
Puissance passive : le mû, **supporte ou non** le déplacement (fonction de ses qualités diverses).
Information : donnée factuelle.
Définition d'une information : organisation des concepts mises à l'œuvre dans cette information; représentation formelle; modèle conceptuel.

Le cheval blanc

Alors, évidemment, même les physiciens naïfs n'iraient pas défendre que dans *nom samouraï*, *samouraï* serait un animé, une personne, un guerrier, un japonais ou bien un noble. Pourquoi donc dans *cheval blanc*, *cheval* serait-il d'emblée un animé, un animal, une monture ou bien un eumétazoaire?



La couleur du *cheval blanc*, en quelques traits



Terme et concept introduits

Surface comme entité

Couleur comme propriété

Terme et concept introduits

Surface comme entité

Cette solution au problème posé ressemble finalement étrangement au *dictionnaire à l'envers* (1992), mais avec beaucoup plus de points d'appui.

Conclusion

La résolution du *cheval blanc* a abouti à
• une solution systémique sans niveau d'analyse
• mélangeant *intension comme **fond*** et *intension comme **forme***
Cette solution changea des habitudes en terme
• informatique (architecture rigide de la version 1)
• tentation de raisonner par classes (*animé etc.*) préconçues plutôt que par observation directe d'un résultat localement non ambigu
La solution est structurale et constructiviste. Elle comporte :

une seule structure - une seule analyse - un seul type de résultat
un seul signe dans lequel on n'opère que des sélections de sens

Cette solution emporte
- un moteur de QR susceptible d'effectuer sa recherche dans tout le graphe créé
- une annotation de signification des mots non triviale
- des résolutions de **coréférence**
- **des filtres d'extraction d'information**.

Conclusion générale

Dans ce document, au plan scientifique, nous avons :

- **clarifié** un peu certaines questions de sémantique lexicale, particulièrement en rapport avec SemCo et les catégories, **l'information et la définition,** ceci en refusant la surinterprétation inhérente à l'emploi des catégories.
- montré comment progresser dans la modélisation du dictionnaire formel LDI
- intégré
  - corpus et dictionnaire,
  - éléments compositionnels et syntaxiques,
  - calcul de la référence et interprétation,
  - dénomination naturelle et contenu formel.

Conclusion générale

En terme applicatif, l'ouverture de l'intensionnalité sous deux dimensions, nous a conduit à l'intégration de :

- un moteur à la Google
- l'étiquetage par apprentissage à chaud
- un système de QR à base de sémantique lexicale
- un système endogénéisant certains calculs de coréférence
- la recherche et le traitement des locutions
- l'information et sa définition

Ce document peut fournir un plan de travail (linguistique, graphe, algo) pour de nombreuses années. Il a été rendu possible par la motivation de la HDR.

Pour m'avoir conduit jusqu'ici, je vous remercie.

Non traité

Comment vérifier l'aptitude de la puissance passive à se prêter à une action :
→ voir nom samouraï. *Le principe est le même.*

Liens avec Framenet?
Framenet est également un projet constructiviste. Il peut fournir :
-Certains modèles
-Une aide sur le choix du lexique.

Peut-on traiter de : *l'avocat mange*
Oui. Après avoir traité de la faim, de la satiété, du besoin, de l'importance biologique de manger, de la disparition, de la survivance etc.
C'est précisément possible mais ne prend une forme :
Manger → (implique)
Mais prend une forme
Manger *est co-occurrent de XXX*, comme le nombre de radios vendu dans les années 1920/30 était directement en corrélation avec le nombre de fous dans les villages.