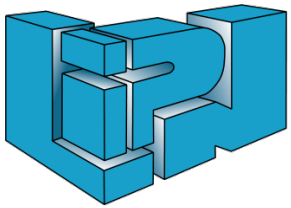
Analyse de concepts formels pour la construction d'ontologies à partir de textes

la question du corpus

Thibault Mondary, Sylvie Després

Laboratoire d'Informatique de Paris-Nord



27 janvier 2009

Introduction

- Cimiano *et al.* (2005) puis Bendaoud *et al.* (2007) utilisent l'analyse de concepts formels pour construire une hiérarchie de concepts à partir de textes
- Nous voulons plutôt aider l'expert à fabriquer une hiérarchie de concepts dénotés (une ontologie lisible par l'humain et utilisable par la machine)
- Dans le cadre du droit international du travail, les résultats sont décevants
- Sur un autre corpus, ils semblent plus utilisables par l'expert.

Objectifs

- Peut-on exhiber des indicateurs de l'interprétabilité (quantitative) du treillis en fonction du corpus ?
- Comment améliorer l'utilisabilité du treillis dans le cadre du travail de conceptualisation par l'expert ?

Définitions

Qu'appelons-nous "ontologie" ?

D'après (Studer, 98 ; Grüber, 93) : *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*

La conceptualisation, c'est :

- L'identification des entités représentatives d'un domaine
- La construction du modèle conceptuel
- Une étape difficile au coeur du processus de construction d'ontologies

Pourquoi partir de textes ?

- Ils sont porteurs de connaissances stabilisées et partagées
- Ils sont plus accessibles que les experts
- Ils facilitent le retour au texte

Analyse de Concepts Formels (ACF)

Principes

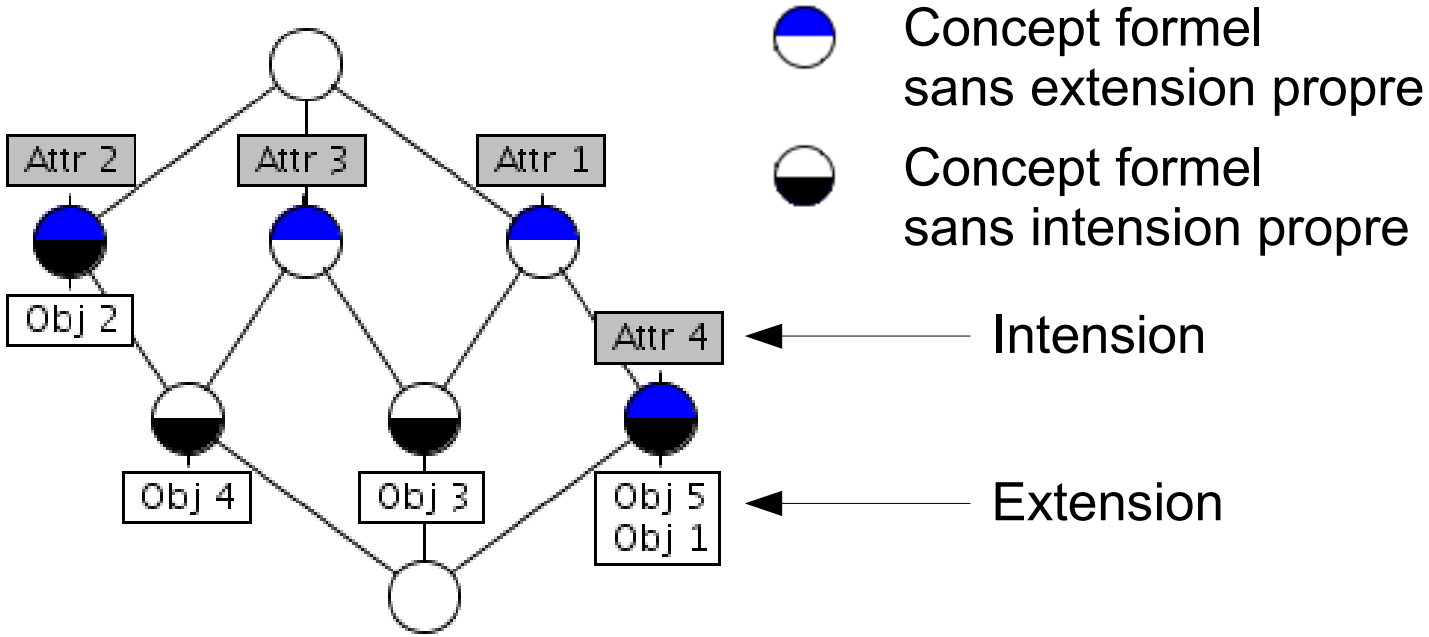
- Méthode de classification symbolique proposée par Ganter & Wille (1999)
- Découverte de tous les regroupements possibles d'éléments ayant des traits en commun
- Construction d'un treillis de concepts à partir d'un contexte formel
- Un contexte formel est un triplet $\mathbb{K} = (G, M, I)$
- Opérateur dual :
  - $\forall A \subseteq G, A' = \{m \in M | \forall g \in A, (g, m) \in I\}$
  - $\forall B \subseteq M, B' = \{g \in G | \forall m \in B, (g, m) \in I\}$
- Un concept formel : (A, B) tel que $A' = B$ et $B' = A$

ACF (2)

	Attr 1	Attr 2	Attr 3	Attr 4
Obj 1	×			×
Obj 2		×		
Obj 3	×		×	
Obj 4		×	×	
Obj 5	×			×

ACF (2)

	Attr 1	Attr 2	Attr 3	Attr 4
Obj 1	×			×
Obj 2		×		
Obj 3	×		×	
Obj 4		×	×	
Obj 5	×			×



ACF et textes

Principes

- Popularisé par Cimiano *et al.* (2005) puis par Bendaoud *et al.* (2007)
- Utilisation de la structure syntaxique des phrases (dans la lignée de Hindle (1990))
- Se concentre sur les relations `sujet/verbe` et `complément/verbe` pour construire le contexte formel

Exemple : "*Une grève paralyse l'entreprise. Les travailleurs refusent le travail dominical. Les dirigeants refusent la grève.*

ACF et textes

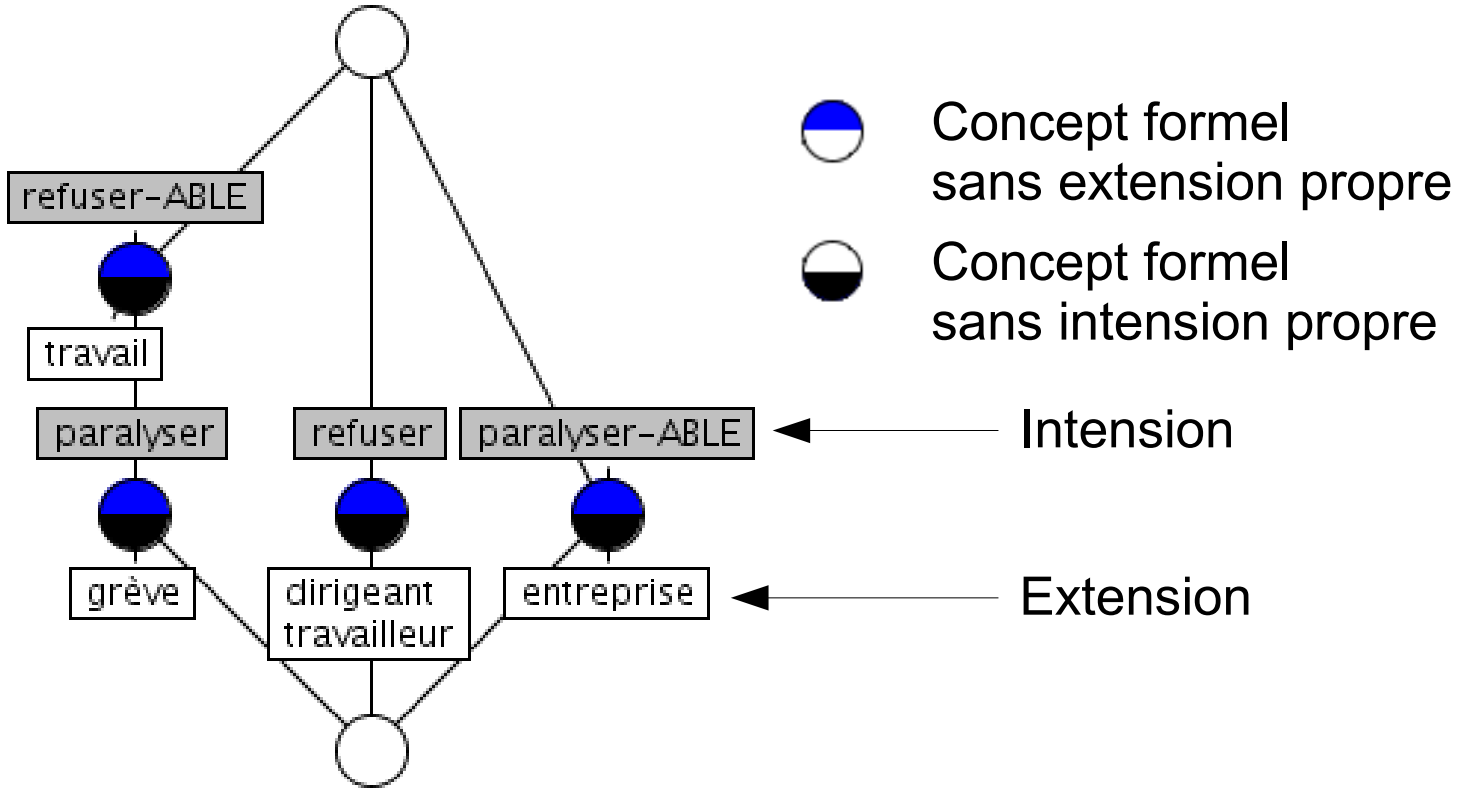
Principes

- Popularisé par Cimiano *et al.* (2005) puis par Bendaoud *et al.* (2007)
- Utilisation de la structure syntaxique des phrases (dans la lignée de Hindle (1990))
- Se concentre sur les relations `sujet/verbe` et `complément/verbe` pour construire le contexte formel

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 (2)



Corpus

Deux corpus de taille similaire (environ 3600 mots)

- C87_C98, provenance `http://www.ilo.org/ilolex` :
  - Convention C87 sur la liberté syndicale et la protection du droit syndical, 1948
  - Convention C98 sur le droit d'organisation et de négociation collective, 1949
- Recettes de cuisine, provenance `http://cuisineaz.com`
  - Une trentaine de recettes
  - Une dizaine de lignes par recette

Un corpus plus volumineux (environ 100 000 mots)

- Les deux conventions C87, C98 et la jurisprudence associée

Corpus

Deux corpus de taille similaire (environ 3600 mots)

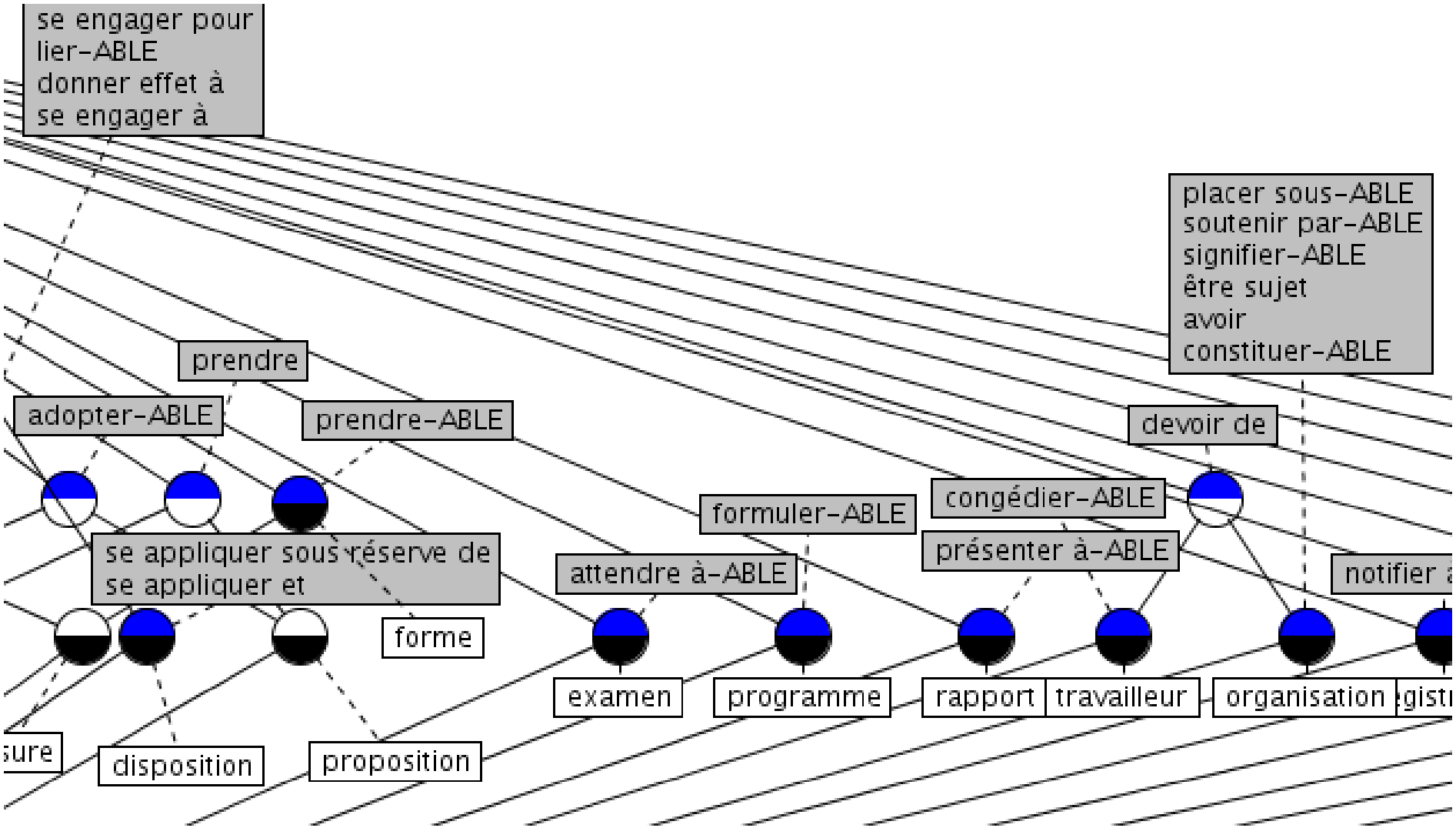
- C87_C98, provenance `http://www.ilo.org/ilolex` :
  - Convention C87 sur la liberté syndicale et la protection du droit syndical, 1948
  - Convention C98 sur le droit d'organisation et de négociation collective, 1949
- Recettes de cuisine, provenance `http://cuisineaz.com`
  - Une trentaine de recettes
  - Une dizaine de lignes par recette

Un corpus plus volumineux (environ 100 000 mots)

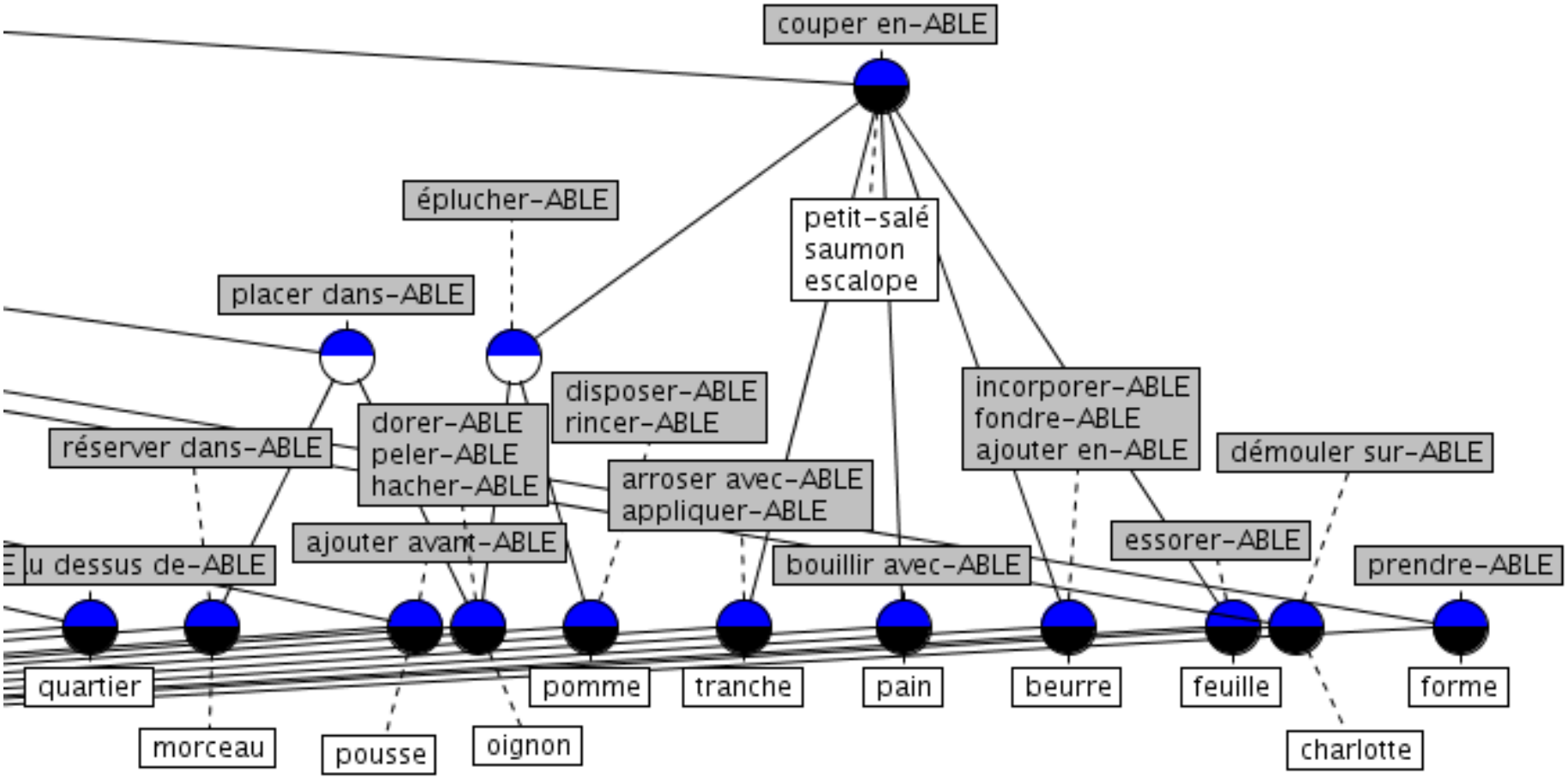
- Les deux conventions C87, C98 et la jurisprudence associée

Résultats...

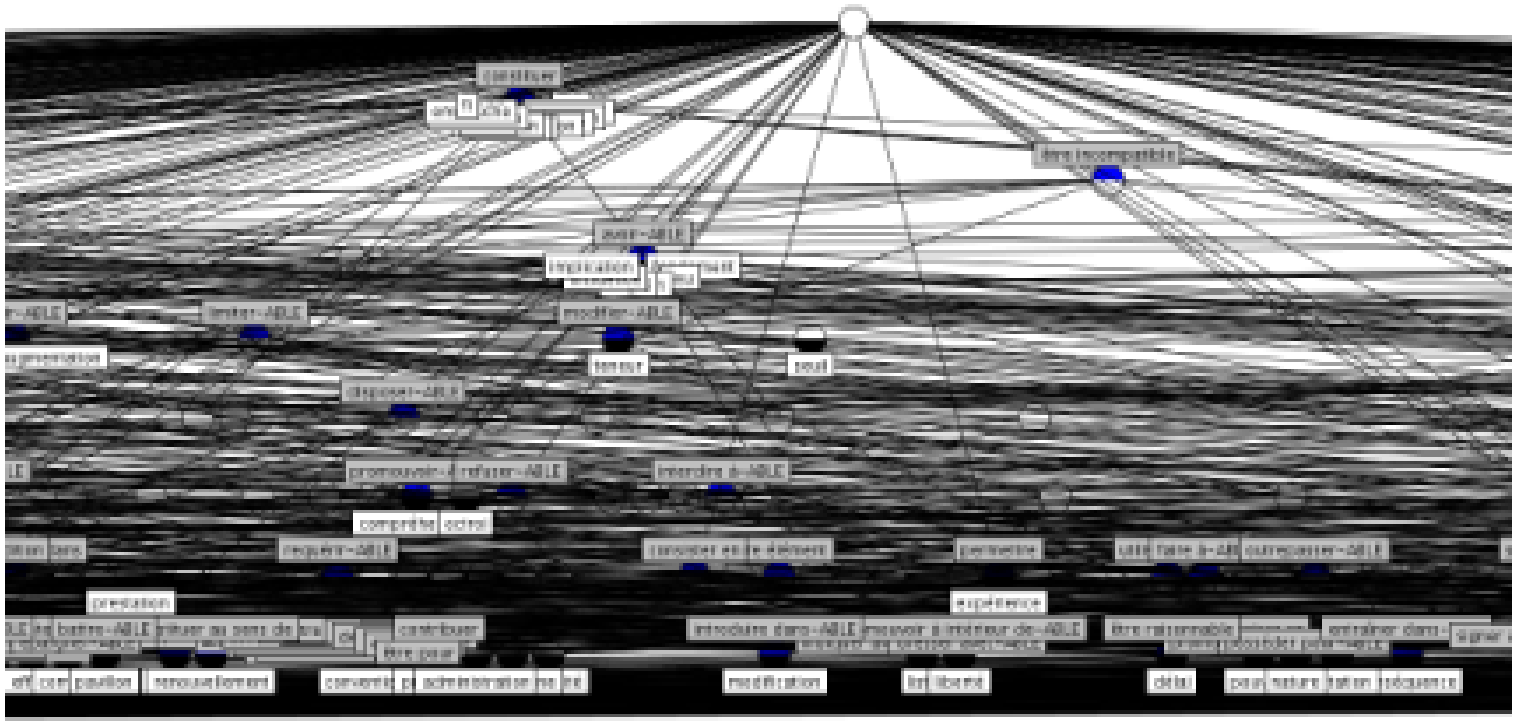
Treillis pour les conventions



Treillis pour les recettes



Treillis conventions+jurisprudences



Indicateurs d'interprétabilité ?

	C87_C98	C87_C98_Jurisp	Recettes
Domaine	liberté syndicale	liberté syndicale	cuisine
Nombre de mots	3605	99966	3644
Nombre de phrases	158	4533	299
Phrase la plus longue (en mots)	237	292	58
Nombre moyen de mots par phrase	22,82	22,05	12,18
Nombre de verbes	430	14137	666
Dont modaux	36	1195	6

Indicateurs d'interprétabilité ? (2)

	C87_C98	C87_C98_Jurisp	Recettes
Taille du vocabulaire lemmatisé	868	5858	1198
Fréquence moyenne du voc. lemmatisé	4,15	17,06	3,04
Taille du vocabulaire fléchi	1102	9686	1422
Fréquence moyenne du voc. fléchi	3,27	10,32	2,56
Nombre de candidats termes	122	4064	199
Nombre de termes retenus	109	3297	162

Indicateurs d'interprétabilité ? (2)

	C87_C98	C87_C98_Jurisp	Recettes
Taille du vocabulaire lemmatisé	868	5858	1198
Fréquence moyenne du voc. lemmatisé	4,15	17,06	3,04
Taille du vocabulaire fléchi	1102	9686	1422
Fréquence moyenne du voc. fléchi	3,27	10,32	2,56
Nombre de candidats termes	122	4064	199
Nombre de termes retenus	109	3297	162

Comment améliorer l'interprétabilité ?

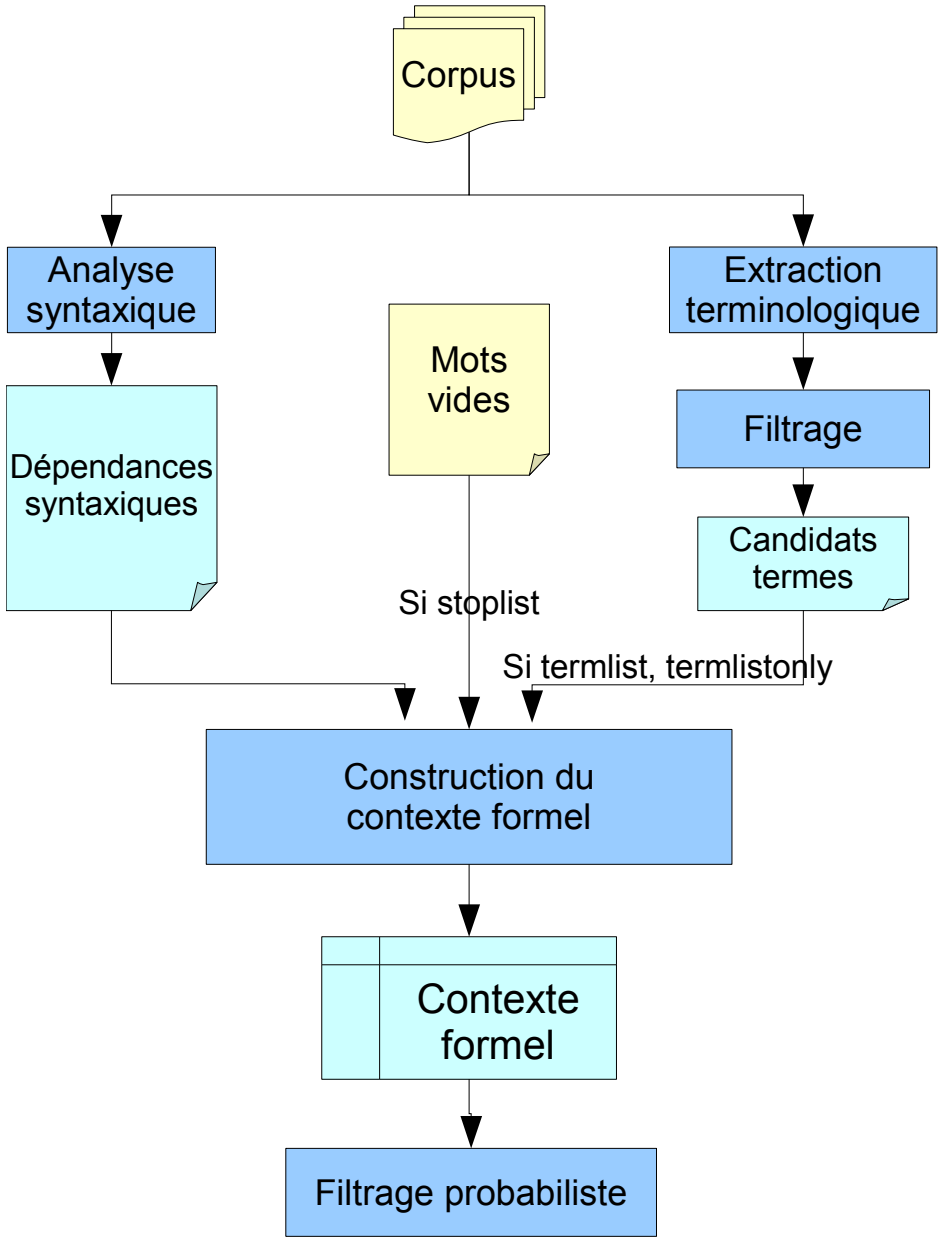
Indicateurs d'interprétabilité ? (2)

	C87_C98	C87_C98_Jurisp	Recettes
Taille du vocabulaire lemmatisé	868	5858	1198
Fréquence moyenne du voc. lemmatisé	4,15	17,06	3,04
Taille du vocabulaire fléchi	1102	9686	1422
Fréquence moyenne du voc. fléchi	3,27	10,32	2,56
Nombre de candidats termes	122	4064	199
Nombre de termes retenus	109	3297	162

Comment améliorer l'interprétabilité ?

Idée : mieux sélectionner les objets du contexte formel

Système


Corpus
Analyse syntaxique
Extraction terminologique
Mots vides
Dépendances syntaxiques
Filtrage
Candidats termes
Si stoplist
Si termlist, termlistonly
Construction du contexte formel
Contexte formel
Filtrage probabiliste

Système (2) : TAL

Extraction de termes : YaTeA (Sophie Aubin, 2006)

- Travaille sur le français et l'anglais
- Extrait des groupes nominaux candidats-termes
- Utilise des patrons d'extractions et une désambiguïsation endogène

Analyse syntaxique : Syntex (Bourigault *et al.*, 2005)

- Travaille sur le français
- Fournit une analyse syntaxique en dépendances
- A obtenu de bons résultats lors de la campagne EASY

Construction du contexte formel

- Repérage des relations sujet/verbe, complément d'objet/verbe
- Ajout du suffixe -ABLE aux relations complément/verbe
- Six variantes de construction axées sur la sélection des objets

Système (3) : variantes de construction du contexte

usuelle : objets = têtes lemmatisées des sujets et des compléments d'objet (si ce sont des noms) ; attributs = verbes associés (+prépositions).

stoplist : Idem, à condition que les candidats n'appartiennent pas à une liste de mots vides fournie par l'utilisateur et ne contiennent pas de chiffre.

termlist : Idem à la méthode usuelle, en tentant d'apparier les candidats objets avec une liste de termes lemmatisés fournie en entrée par l'utilisateur. Si l'appariement échoue, inclure seulement la tête.

termlist+stoplist : ...

termlistonly : N'inclure dans le contexte que les sujets ou compléments d'objet qui s'apparient avec un terme de la liste fournie en entrée. Les verbes associés aux sujets ou compléments sont inclus dans le contexte comme attributs.

termlistonly+stoplist : Comme ci-dessus, mais filtrer les verbes selon une liste de mots vides.

Système (4) : Filtrages

Au niveau des candidats-termes

- Conseillé pour les méthodes `termlist/termlistonly`
- Un filtrage par patrons
- Des patrons existent (des nombres, “article” pour les conventions...)
- Ils doivent être adaptés au corpus (*a priori* moins coûteux que le filtrage exhaustif)

Au niveau des couples du contexte formel

- Dans la lignée de Cimiano *et al.* (2005)
- Calcul de la probabilité conditionnelle de l'objet *obj* sachant l'attribut *attr* pour chaque couple $(obj, attr)$ du contexte formel : $P(obj|attr) = f(obj, attr)/f(attr)$
- Problème : fixer le seuil

Avantages/Inconvénients : méthode `termlist`

- Les objets (du contexte) sont plus "parlant" si c'est possible
- Tous les objets de la méthode usuelle sont conservés

- Diminution des regroupements
- Nécessité d'un filtrage

Avantages/Inconvénients : méthode `termlistonly`

- Les objets (du contexte) sont plus "parlant"
- Dans l'idéal, le treillis ne contient plus que des concepts utiles

- Diminution drastique des regroupements
- Importance du filtrage

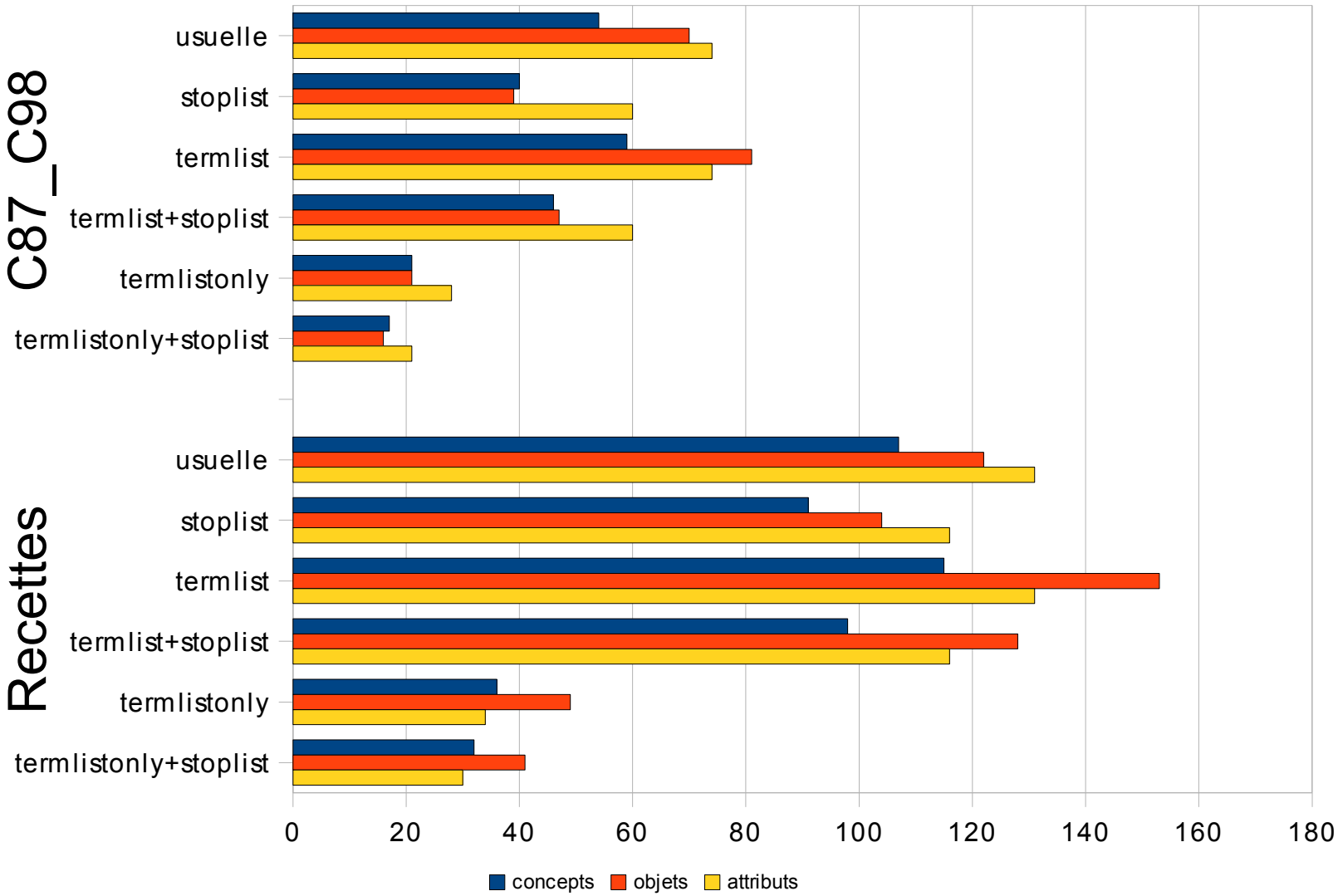
Avantages/Inconvénients : méthode `termlistonly`

- Les objets (du contexte) sont plus "parlant"
- Dans l'idéal, le treillis ne contient plus que des concepts utiles

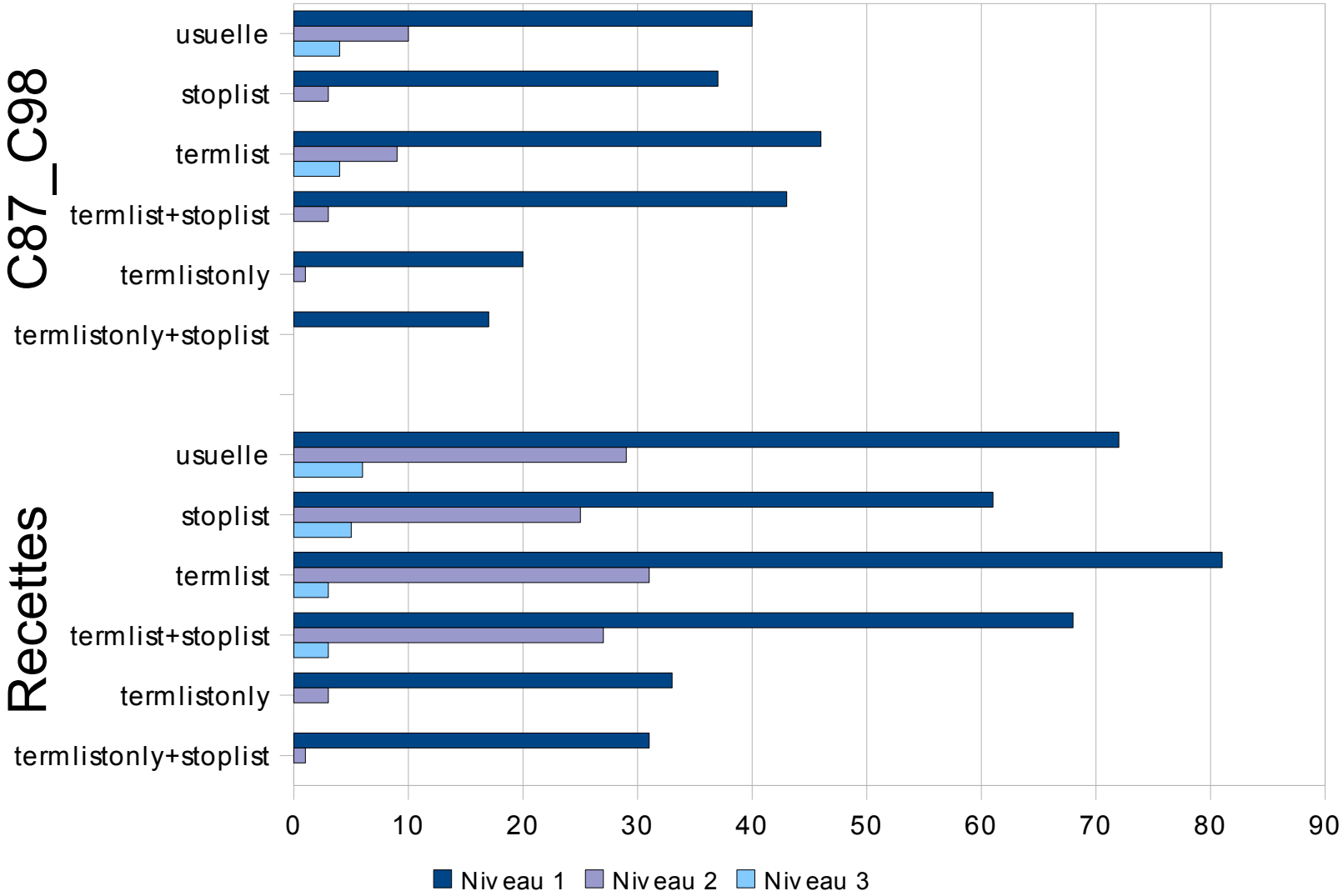
- Diminution drastique des regroupements
- Importance du filtrage

Ce que cela donne au niveau quantitatif

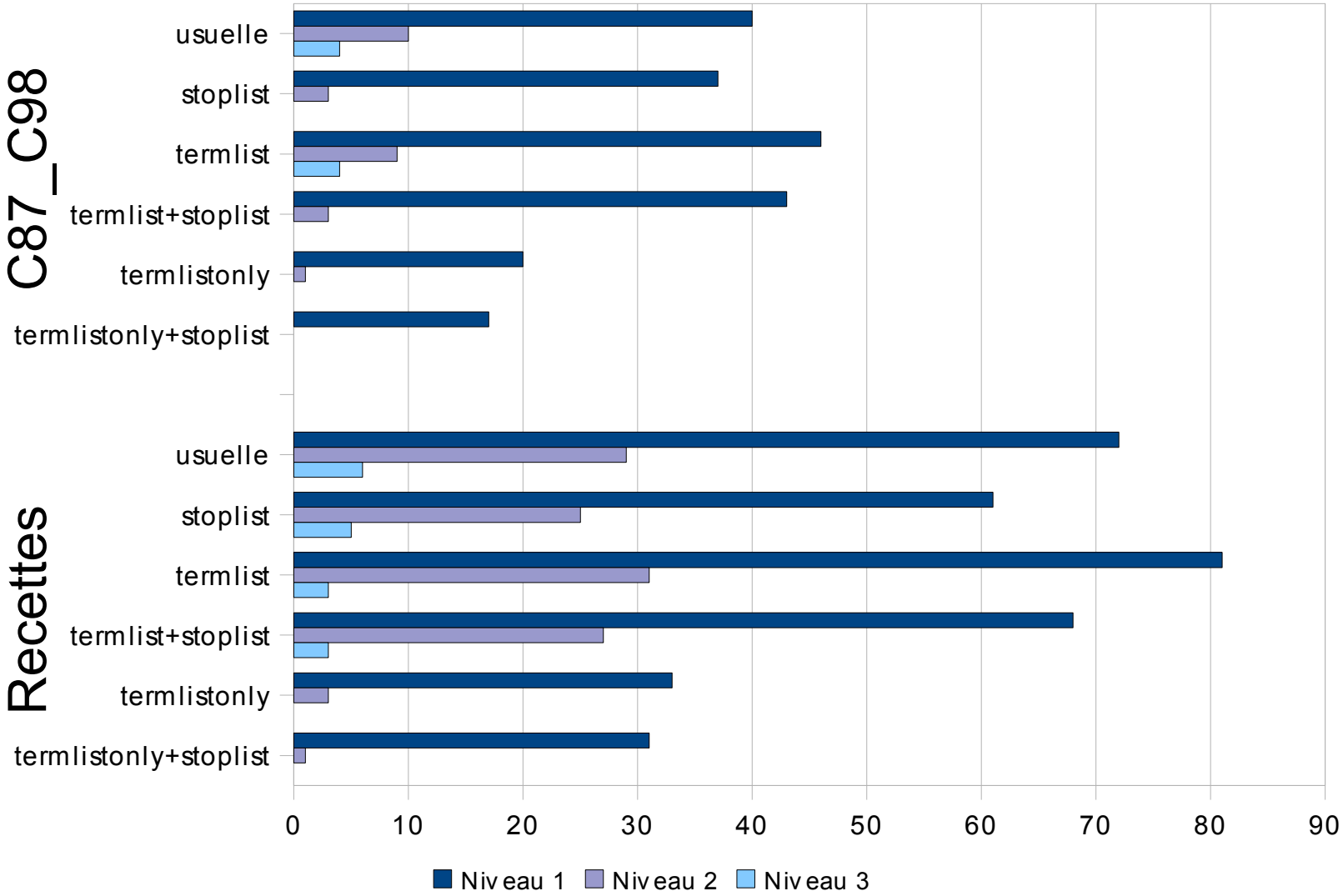
Objets, attributs, concepts



Concepts de niveau 2 et 3

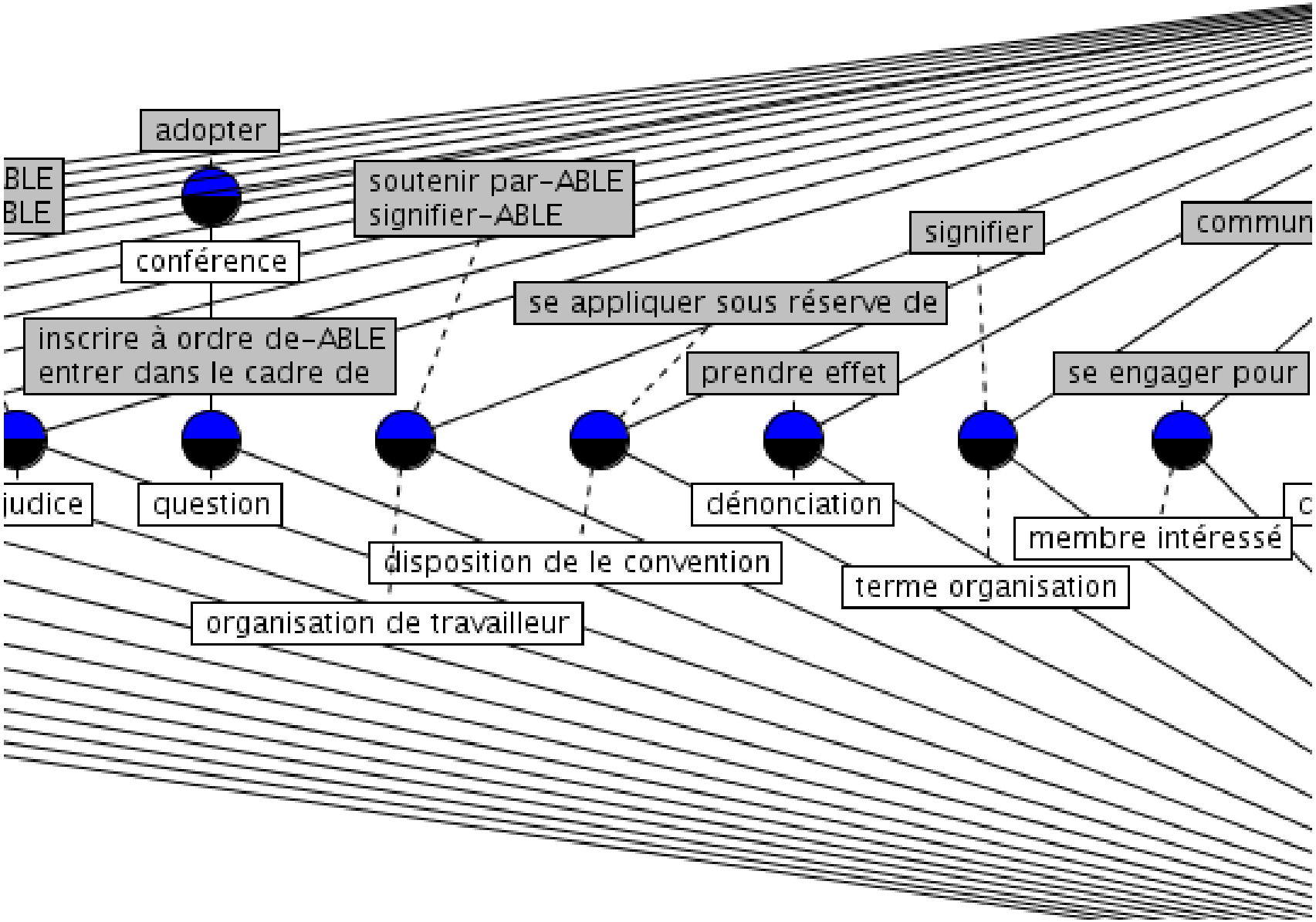


Concepts de niveau 2 et 3



Et au niveau du treillis ?

Treillis pour les conventions, méthode termlist



Conclusion

- L'approche syntaxique nécessite de prendre en compte la "qualité" du corpus
- La taille des phrases et le nombre de verbes de modalité semblent être des indicateurs de l'interprétabilité du treillis par l'expert
- Les variantes `termlist/termlistonly` sont un moyen d'augmenter l'interpretabilité du treillis

Des problèmes

- Mesurer l'interprétabilité de manière qualitative
- Fixer le seuil de filtrage
- Changer de contexte dans les dépendances syntaxiques pour gérer la modalité
- Présenter les données à l'expert (sous forme de treillis ? de listes de concepts ?)

Bendaoud R., Rouane Hacene M., Toussaint Y., Delecroix B. & Napoli A. (2007). Construction d'une ontologie à partir d'un corpus de textes avec l'acf. In F. Trichet, Ed., *Actes des 18eme journées francophones d'ingénierie des connaissances (IC2007)* : Cépaduès.

Bourigault D., Fabre C., Frérot C., Jacques M.-P. & Ozdowska S. (2005). Syntex, analyseur syntaxique de corpus. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, Dourdan, France.

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

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

Hindle D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, p. 268–275, Morristown, NJ, USA : Association for Computational Linguistics.

Sophie Aubin T. H. (2006). Improving term extraction with terminological resources. In T. Salakoski, F. Ginter, S. Pyysalo & T. Pahikkala, Eds., *Advances in Natural Language Processing 5th International Conference on NLP, FinTAL 2006*, p. 380–387, Turku, Finland : Springer.