

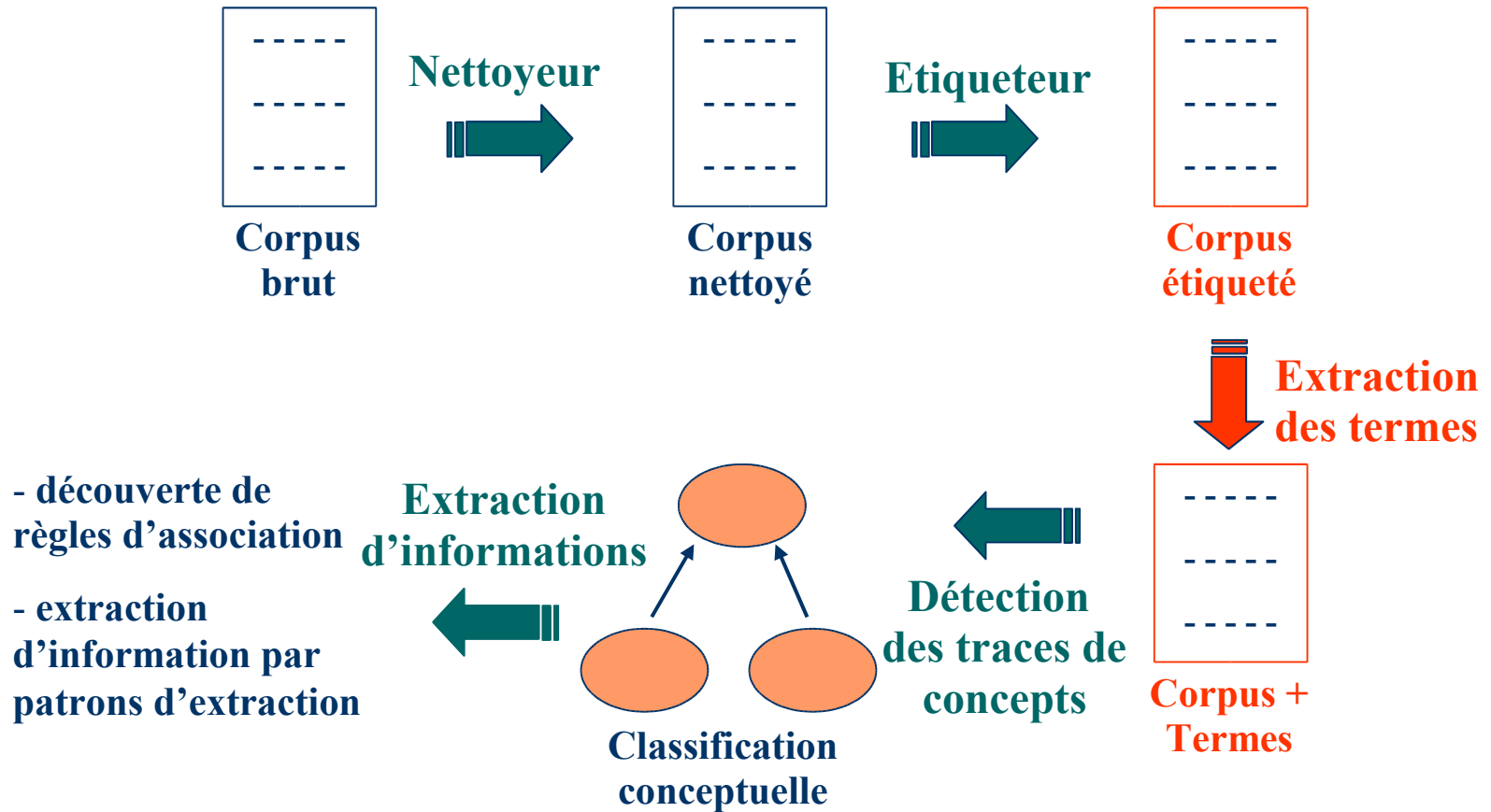
Extraction de la terminologie

Mathieu Roche

Cours Fouille de Données

février 2007

Introduction



Introduction

- **Une définition de la terminologie :**

« Langue particulière que se fait chaque auteur »,
Emile Littré (1876)



Les termes

- **Terme** = groupe de mots ayant des propriétés syntaxiques + trace linguistique de concepts pour une tâche en cours.
 - Exemples, les candidats termes "**intelligence artificielle**" et "**génie logiciel**" sont des termes.
 - Exemple, le candidat terme "**chalon sur saône**" est-il un terme ?

Pourquoi étudier les termes ?

- Importance de la caractérisation des termes
 - Exemple : traduction automatique
 - Constitution européenne, article III-10 :

The right to vote and to stand as a candidate in elections...

Le droit de vote et d'éligibilité aux élections ...

Quelques systèmes d'extraction de la terminologie

Systèmes	linguistiques	statistiques	références
TERMINO	X		[David et Plante 1990]
LEXTER	X		[Bourigault 1993]
FASTR	X		[Jacquemin 1996]
INTEX	X		[Silberztein1994 ; Ibekwe-SanJuan 2001]
ANA		X	[Enguehard 1993]
MANTEX		X	[Frath <i>et al.</i> 2000]
XTRACT	X	X	[Smadja 1993]
ACABIT	X	X	[Daille 1994]
CLARIT	X	X	[Evans et Zhai 1996]
TERMIGHT	X	X	[Daga et Church 1997]
SYNTEX	X	X	[Bourigault et Fabre 2000]
C/NC VALUE	X	X	[Frantzi <i>et al.</i> 2000]
WASPBENCH	X	X	[Kilgariff et Tugwel 2001]
FIPS	X	X	[Nerima <i>et al.</i> 2003]
ESATEC	X	X	[Biskri <i>et al.</i> 2004]
EXIT	X	X	[Roche <i>et al.</i> 2004]

Itératif

Coopératif

LEXTER [Bourigault, 93 ; Jacquemin et Bourigault, 99] (1/8)

- Méthode linguistique
- Trois étapes :
 - Extraction des groupes nominaux maximaux
 - Décomposition des groupes nominaux maximaux
 - Présentation des résultats sous forme d'un réseau sémantique

LEXTER : 1^{ère} étape (2/8)

- Extraction des groupes nominaux maximaux
 - L'idée qui est à la base de la conception de LEXTER est celle de repérage de frontière.
 - Le principe de base est donc de découper le texte en repérant ces frontières potentielles entre lesquelles on isole des syntagmes nominaux susceptibles d'être des occurrences de termes.

LEXTER : 1^{ère} étape (3/8)

- Extraction des groupes nominaux maximaux
 - Les règles de découpage décrivent des marqueurs de frontière sous la forme de patrons morpho-syntaxiques; par exemple : verbe, conjonction, préposition + adjectif possessif, etc.
 - Les données d'entrée du module chargé d'effectuer le découpage sont uniquement des informations morphologiques associées à chaque mot du texte : catégorie grammaticale, traits morphologiques (en particulier genre et nombre), forme lemmatisée.

LEXTER : 1^{ère} étape (4/8)

- Extraction des groupes nominaux maximaux

Texte initial (étiqueté)

le circuit d'aspersion de l'enceinte de confinement assure le maintien de sa température nominale de fonctionnement après une augmentation de pression.

Groupes nominaux maximaux

- circuit d'aspersion de l'enceinte de confinement
- maintien
- température nominale de fonctionnement
- augmentation de pression

Verbe --> coupe
préposition + adj. possessif → coupe
préposition + art. indéfini → coupe

LEXTER : 2^{ème} étape (5/8)

- Décomposition des groupes nominaux maximaux
 - Hypothèse : tout terme complexe est composé d'une tête et d'une expansion.

LEXTER : 2^{ème} étape (6/8)

- Décomposition des groupes nominaux maximaux
 - Deux règles classiques de décomposition
 - nom1 adjectif :
 - Tête : nom1
 - Expansion : adjectif
 - nom1 de nom2 :
 - Tête : nom1
 - Expansion : nom2 (de)

LEXTER : 2^{ème} étape (7/8)

- Problème d'ambiguïtés
 - Par exemples, les groupes nominaux de type Nom1 de Nom2 Adjectif (corps français) : centre de tourisme équestre
 - Problème de rattachement s'il y a absence d'informations sur le genre ou le nombre
 - Deux types de décompositions pour «centre de tourisme équestre»

Tête : centre

Expansion : tourisme équestre

Tête : centre de tourisme

Expansion : équestre

Groupe non ambigu également trouvé dans le corpus --> 1^{ère} décomposition retenue

LEXTER : 3^{ème} étape (8/8)

- Présentation des résultats sous forme d'un réseau sémantique

FASTR [Jacquemin, 96] (1/2)

- **Entrée** : termes de référence (congé de formation)
- **Sortie** : termes variants (congé annuel de formation).
- 3 types de règles (linguistiques) :

<i>coordination</i>	association rule --> association and classification rules
<i>insertions</i>	MRI image --> MRI brain image
<i>permutation</i>	knowledge discovery --> discovery of knowledge

FASTR (2/2)

- **Remarque** : dans certains cas, nécessité de considérer une fenêtre plus grande [Ville-Ometz *et al.* 2004].

Exemple : **thymus gland** --> **thymus** and adrenal **gland**

contexte : **rat** **thymus** and adrenal **gland**

ANA (Apprentissage Naturel Automatique)

[Enguehard, 93 ; Enguehard, 01] (1/8)

- Méthode numérique
- Méthode incrémentale
- Deux étapes :
 - Module « Familiarisation »
 - Module « Découverte »

ANA : 1^{ère} étape (2/8)

- Module « Familiarisation » : extraction de connaissances dans les textes sous forme de quatre liste.
- 1^{ère} liste : les **mots fonctionnels** : articles, pronoms, adverbes. Liste établie statistiquement.
"a", "alors", "après", "au", "auraient", "aussi", "autre", "avait", "avant", "avec", "avoir", "beaucoup", "c", "car", "ce", "cela", "celles", "certain", "ces", "cette", "ceux", "chacun", "chaque", "comme", "comment", "d", "dans", "de", "déjà", "des", "dirais", "dire", "dit", "donc", "du", "elle", "en", "encore", "est", "et", "était« , etc.

ANA : 1^{ère} étape (3/8)

- Module « Familiarisation »
 - 2^{ème} liste : les **mots fortement liés** : variation morphologiques de certains mots fonctionnels. Par exemple, « de la », « est en », « est le ».
 - 3^{ème} liste : les **mots de schémas** : mots fonctionnels structurant les groupes de mots. Par exemple, « de », « de la », « des », « du », « en », etc.
 - 4^{ème} liste : les **bootstrap** : quelques termes du domaine.

ANA : 2^{ème} étape (4/8)

- Module « Découverte »
 - ANA consiste à enrichir, de manière incrémentale, les termes du bootstrap de trois manières différentes.
 - Exemples en utilisant les termes du bootstrap suivant : {**automate, centrale, circuit, cœur, cuve, fréquence, gaz, rédacteur, structures, tubes, vibration, vitesse**}

ANA : 2^{ème} étape (5/8)

- Module « Découverte »

Bootstrap = {**automate, centrale, circuit, cœur, cuve, fréquence, gaz, rédacteur, structures, tubes, vibration, vitesse**}

- 1^{er} cas : les cooccurrences extraites dans le corpus possèdent deux termes du bootstrap. Exemple,
 - **réacteur dont le cœur (1)**
 - **coeur de ce réacteur (1)**
 - **cœur du réacteur (3)** → **Nouveau terme**
 - **coeur le réacteur (1)**

ANA : 2^{ème} étape (6/8)

- Module « Découverte »

Bootstrap = {**automate, centrale, circuit, cœur, cuve, fréquence, gaz, rédacteur, structures, tubes, vibration, vitesse**}

- 2^{ème} cas : les cooccurrences extraites dans le corpus possèdent un terme du bootstrap, un mot de schéma et un mot quelconque. Exemple,

- **cuve** du **barillet** (3)

 **Nouveau terme**

ANA : 2^{ème} étape (7/8)

- Module « Découverte »

Bootstrap = {**automate, centrale, circuit, cœur, cuve, fréquence, gaz, rédacteur, structures, tubes, vibration, vitesse**}

- 3^{ème} cas : les cooccurrences extraites dans le corpus possèdent un seul terme du bootstrap et aucun mot de schéma. Le nouveau terme sera une chaîne de caractères composée du terme et d'un autre mot (non fonctionnel). Exemple,

- ici ensuite les **structures internes**

- sans les **structures** acier

- conception des **structures internes**

- assembler les **structures** externes

- démonter les **structures internes**

Nouveau terme

ANA : 2^{ème} étape (8/8)

- Module « Découverte »
 - Les nouveaux termes respectant les trois cas décrits sont rajoutés au bootstrap pour les prochaines itérations.
 - Lorsque aucun nouveau terme n'est repéré, le traitement prend fin.

ACABIT (Automatic Corpus-based Acquisition of Binary Terms) [Daille, 94; Daille, 96] (1/9)

- Méthode mixte.
- Termes proposés sous forme lemmatisée.
- **Deux étapes :**
 - Extraction des termes simples respectant des schémas syntaxiques simples puis extraction des termes plus complexes.
 - Classement des termes selon une mesure statistique.

ACABIT : 1^{ère} étape (2/9)

Extraction des candidats termes

- Déterminer des termes de base :
 - Nom Adjectif --> **connaissance informatique**
 - Nom1 à (Déterminant) Nom2 --> **aide à domicile**
 - Nom1 de (Déterminant) Nom2 --> **contrat de travail**
 - Nom1 Préposition Nom2 --> **vente par téléphone**
 - Nom1 Nom2 --> **machine outil**

ACABIT : 1^{ère} étape (3/9)

Extraction des candidats termes

- Définition d'opérations afin de décomposer les termes complexes en termes de base
 - Combinaisons de termes par la coordination :
 - “Nom1 de Nom3” + “Nom2 de Nom3” --> “Nom1 et Nom2 de Nom3”
 - **envoi de courrier + réception de courrier**
--> **envoi et réception de courrier**
 - Combinaisons de termes par la surcomposition :
 - “Nom1 Préposition1 Nom2” + “Nom1 Préposition2 Nom3”
--> “Nom1 Préposition1 Nom2 Préposition2 Nom3”
professeur de musique + professeur à domicile
--> **professeur de musique à domicile**

ACABIT : 1^{ère} étape (4/9)

Extraction des candidats termes

- Définition d'opérations afin de décomposer les termes complexes en termes de base
 - Modifications syntaxiques des termes par l'addition d'un modifieur adjectival ou adverbial :
 - Modifieur adjectival :
 - “Nom1 Préposition Nom2” --> “Nom1 Adjectif Préposition Nom2”
 - **assistance par téléphone** --> **assistance technique par téléphone**
 - Modifieur adverbial :
 - “Nom Adjectif” --> “Nom Préposition Adjectif”
 - **anglais perfectible** --> **anglais facilement perfectible**

ACABIT : 1^{ère} étape (5/9)

Extraction des candidats termes

- But : compter les couples de lemmes qui respectent les patrons syntaxiques

Patron « Nom1 (Préposition (Déterminant)) Nom2 »	Séquences extraites à partir du corpus	Nbre d'occ.
(centre, formation)	centre de formation	17
	centre régional de formation (modifieur adjectival)	2
	centre expérimental de formation (modifieur adjectival)	1
		20

ACABIT : 1^{ère} étape (6/9)

Extraction des candidats termes

- envoi et réception de courrier -->
(envoi, courrier) et (réception, courrier)
sont comptabilisés une fois chacun.
- ACABIT parcourt le corpus et compte les couples de mots.

ACABIT : 2^{ème} étape (7/9)

Utilisation de mesures statistiques

- Information Mutuelle [Church et Hanks, 90]

$$IM(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \quad \Rightarrow \quad I(x,y) = \log_2 \frac{nb(x,y)}{nb(x)nb(y)}$$

- Information Mutuelle au Cube [Daille, 94]

$$I^3(x,y) = \log_2 \frac{nb^3(x,y)}{nb(x)nb(y)}$$

ACABIT : 2^{ème} étape (8/9)

Exemples de termes extraits sur un corpus de CVs avec l'Information Mutuelle et l'Information Mutuelle au Cube.

Termes Nom-Prép-Nom avec l'information mutuelle

1. beurre de karité (3)
2. jéjunum de rat (3)
3. puy en velay (3)
4. chalon sur saône (4)
- ...

Termes Nom-Prép-Nom avec l'information mutuelle au cube

1. mise en place (111)
2. traitement de texte (57)
3. tableau de bord (23)
4. contrat de qualification (31)
- ...

ACABIT : 2^{ème} étape (9/9)

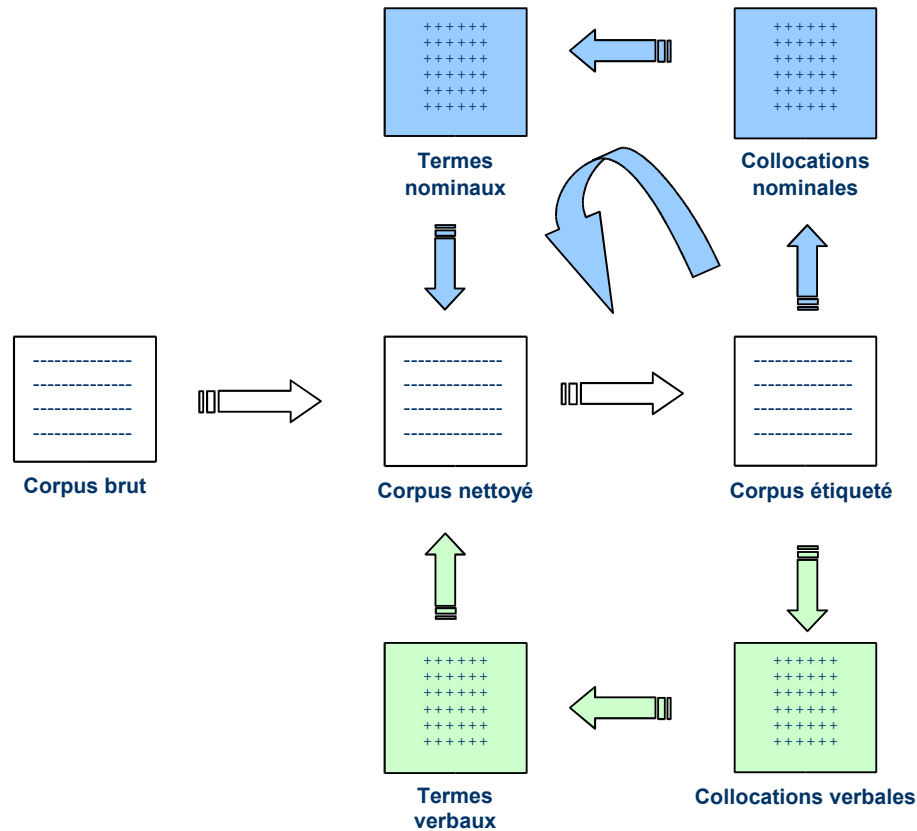
- Rapport de Vraisemblance [Dunning, 93]

	<i>y</i>	<i>y' avec y' ≠ y</i>
<i>x</i>	a	b
<i>x' avec x' ≠ x</i>	c	d

$$\begin{aligned} RV = & a \log(a) + b \log(b) + c \log(c) + d \log(d) - (a+b) \log(a+b) \\ & - (a+c) \log(a+c) - (b+d) \log(b+d) - (c+d) \log(c+d) \\ & + (a+b+c+d) \log(a+b+c+d) \end{aligned}$$

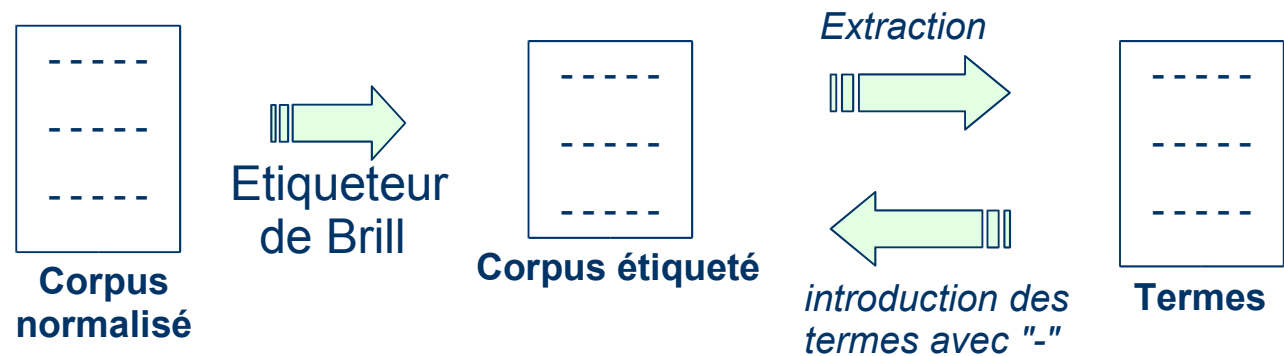
EXIT (EXtraction Itérative de la Terminologie)

[Roche *et al.*, 04] (1/12)



EXIT : Processus itératif (2/12)

- Processus itératif pour extraire les termes nominaux, adjectivaux et adverbiaux.



Exemple :

1^{ère} itération : assistant de gestion

2^{ème} itération : assistant-de-gestion de production

EXIT : Utilisation de mesures statistiques (3/12)

- Information Mutuelle [Church et Hanks, 90]
- Information Mutuelle au Cube [Daille, 94]
- Rapport de Vraisemblance [Dunning, 93]

EXIT : Utilisation de mesures statistiques (4/12)

- Mesure d'Association [Jacquemin, 97] :
 - isobarycentre des valeurs normalisées de l'information mutuelle et du nombre d'occurrences.

EXIT : Utilisation de mesures statistiques (5/12)

- Coefficient de Dice [Smadja, 96]

$$Dice(x,y) = \frac{2P(x,y)}{P(x)+P(y)}$$

$$\Rightarrow D(x,y) = \frac{2 \text{nb}(x,y)}{\text{nb}_{type}(y) \cdot \text{nb}(x) + \text{nb}_{type}(x) \cdot \text{nb}(y)}$$

EXIT : Expérimentations, mesures d'évaluation (6/12)

- **Evaluation des mesures** : Rappel de la notion générale de **précision** et de **rappel**

$$précision = \frac{\text{nombre d'exemples positifs couverts}}{\text{nombre d'exemple couverts}}$$

Une précision de 100% signifie que tous les exemples couverts sont positifs.

$$rappel = \frac{\text{nombre d'exemples positifs couverts}}{\text{nombre d'exemples positifs}}$$

Une couverture de 100% signifie que tous les exemples positifs sont couverts.

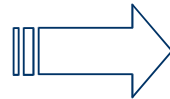
EXIT : Expérimentations, mesures d'évaluation (7/12)

- Evaluation des mesures en terminologie : la précision

$$\textit{précision} = \frac{\text{nombre de candidats termes extraits pertinents}}{\text{nombre de candidats termes extraits}}$$

1. real world
2. neural network
3. frequent itemset
4. remote sensing
5. naive bayes
...

Collocations extraites



1. **real world**
2. **neural network**
3. **frequent itemset**
4. remote sensing
5. **naive bayes**
...

EXIT : Expérimentations, mesures d'évaluation (8/12)

- **Evaluation des mesures en terminologie : la précision**

Les courbes d'élévation (« lift chart ») : variation de la précision en fonction du nombre de termes proposés à l'expert.

EXIT : Expérimentations, mesures d'évaluation (9/12)

- Evaluation des mesures en terminologie : le rappel

$$rappel = \frac{\text{nombre de candidats termes extraits pertinents}}{\text{nombre de candidats termes pertinents}}$$

- ***Impossible à calculer !***

EXIT : Expérimentations, protocole expérimental (10/12)

- Corpus de Fouille de Données, de CV, de Ressources Humaines: **termes pertinents** qui sont traces de concepts.
 - *642 termes expertisés --> corpus de Fouille de Données (en anglais) (FD)*
 - *412 termes expertisés --> corpus de CVs (en français) (CV)*
 - *2960 termes --> corpus des Ressources Humaines (en français) (RH)*

EXIT : Expérimentations : corpus de Fouille de Données, de CV et des Ressources Humaines (11/12)

- Elagage à 3

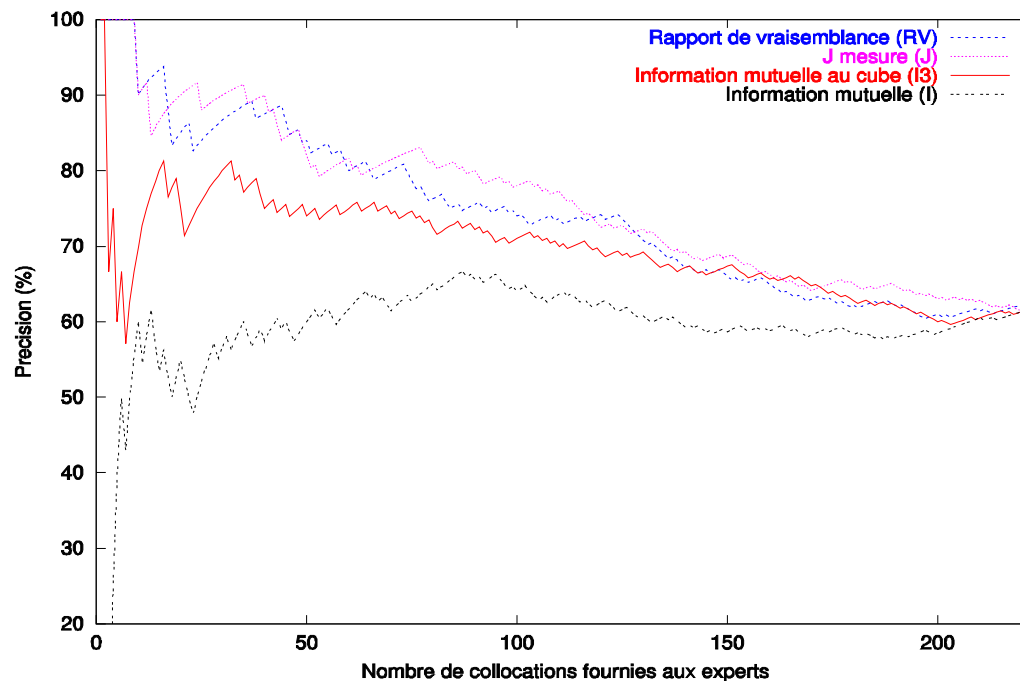
	Nb collocations			Nb collocations après élagage		
	<i>FD</i>	<i>RH</i>	<i>CV</i>	<i>FD</i>	<i>RH</i>	<i>CV</i>
Nom-Prep-Nom	313	4703	3634	7	1268	307
Nom-Nom	2070	98	1781	223	11	162
Adjectif-Nom	2411	1260	1291	176	478	103
Nom-Adjectif	X	5768	3455	X	1628	448

Exemples :

emploi solidarité
action communication
fichier client
service achat
...

EXIT : Expérimentations : corpus de Fouille de Données (*relation Nom-Nom*) (12/12)

- Courbes d'élévation avec quatre mesures.



Conclusion

- **Trois types d'approches pour extraire la terminologie :**
 - linguistique
 - statistique
 - mixte
- **Difficulté : les types de termes extraits peuvent être différents selon les domaines de spécialité** (par exemple, en médecine et en biologie, les termes complexes sont plus pertinents)
 - > **Utilisation de méthodes plus ou moins spécifiques selon les domaines** (exemple, la mesure C/NC-value [Frantzi *et al.*, 00] particulièrement bien adaptée aux domaines de la médecine et de la biologie).