

# **Une chaîne globale de fouille de textes**

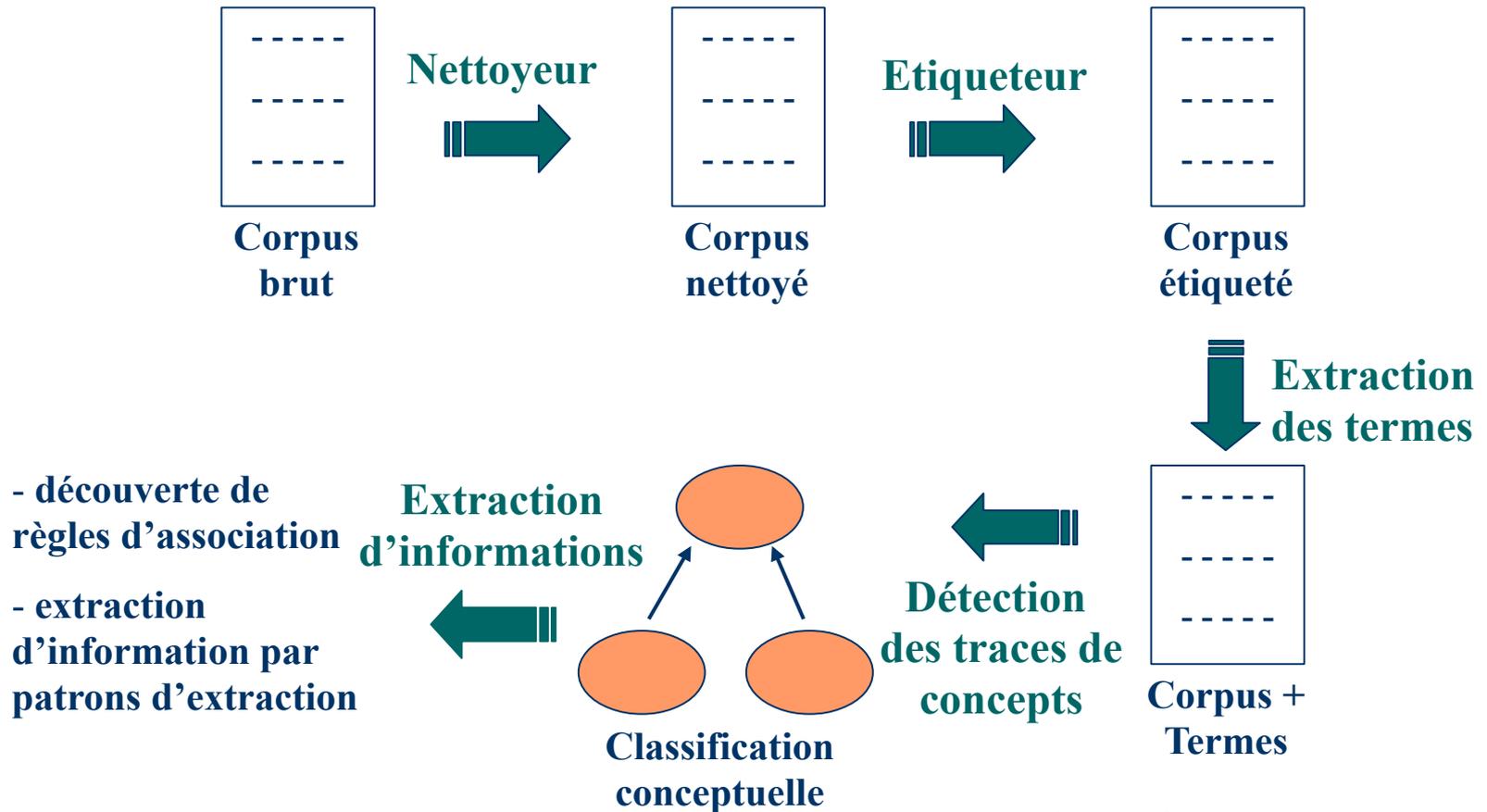
## ***Etiquetage grammatical***

**Mathieu Roche**

**Cours ECDA – M2**

**2016/2017**

# Processus de fouille de textes



# Etape 1 : Le nettoyage

## Exemples de corpus spécialisés :

- Corpus de 100 introductions d'articles en anglais écrits par des auteurs anglophones sur le domaine de la « fouille de données » (369 Ko).
- Corpus de plus de 6000 résumés d'articles en anglais sur la biologie Moléculaire (9424 Ko).
- Corpus en français de plus de 1000 Curriculum Vitæ (VediorBis, 2470 Ko).
- Corpus en français relatif aux Ressources Humaines (PerfomanSe, 3784 Ko).

# Etape 1 : Le nettoyage

- Types de nettoyage :

- Enlever les noms, prénoms, coordonnées, etc. (pour les articles et les CVs)

- Uniformiser les références

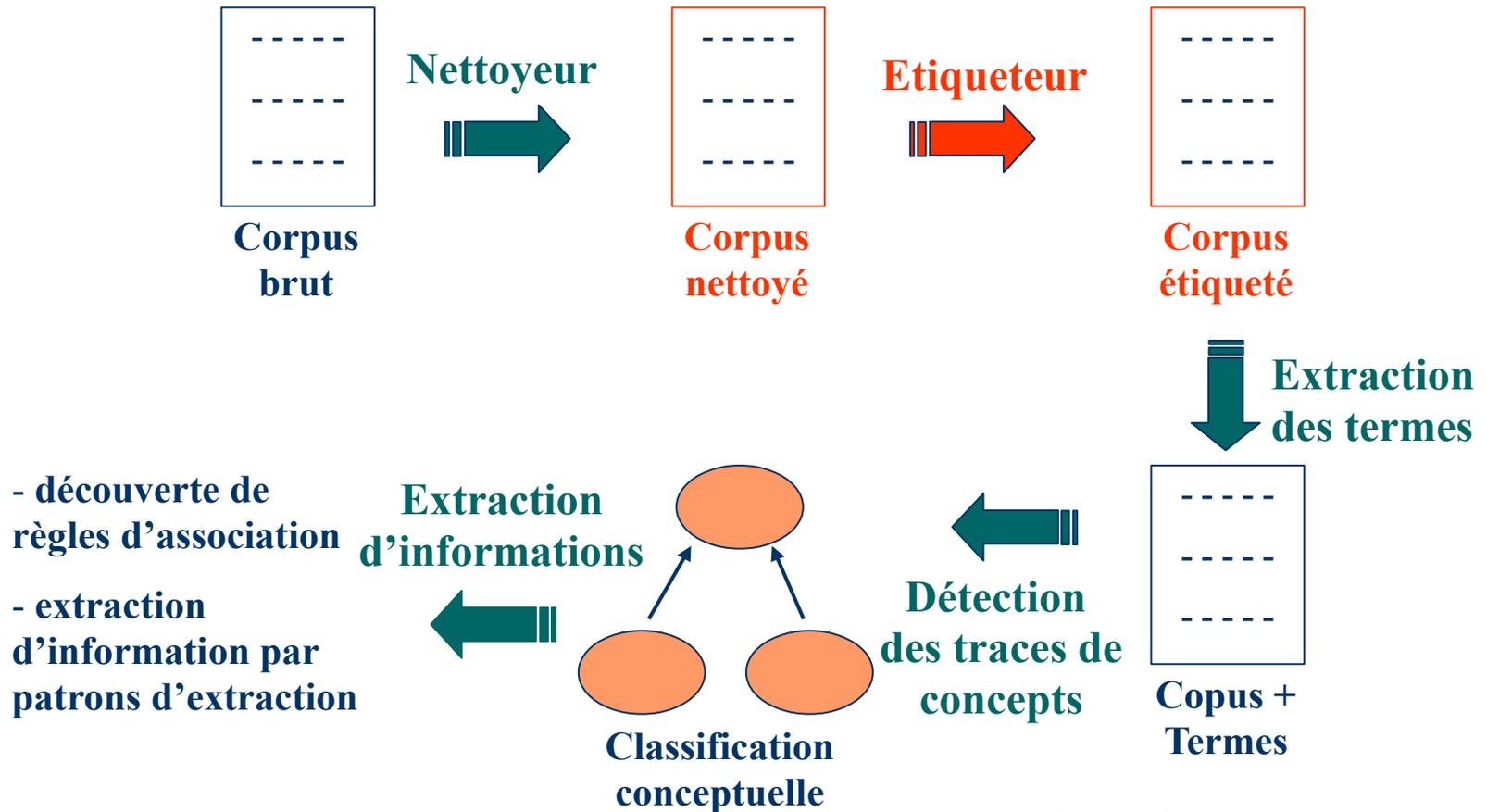
**CORPUS FOUILLE DE DONNEES** : Remplacer ([lettres+année], [numéro], etc.) par « a paper » ou « papers » si ces références sont précédées de la préposition « in », sinon on supprime ces références.

- Généraliser certains noms :

**CORPUS DE BIOLOGIE MOLECULAIRE**

Remplacer : carboxyl-terminal, carboxyl-termini, C00H-terminal, C02H-terminal, etc. par C-term.

# Processus de fouille de textes



## Etape 2 : Etiquetage

Mais pour des  
personnes très  
spontanées ...



**Étiqueteur  
de Brill**

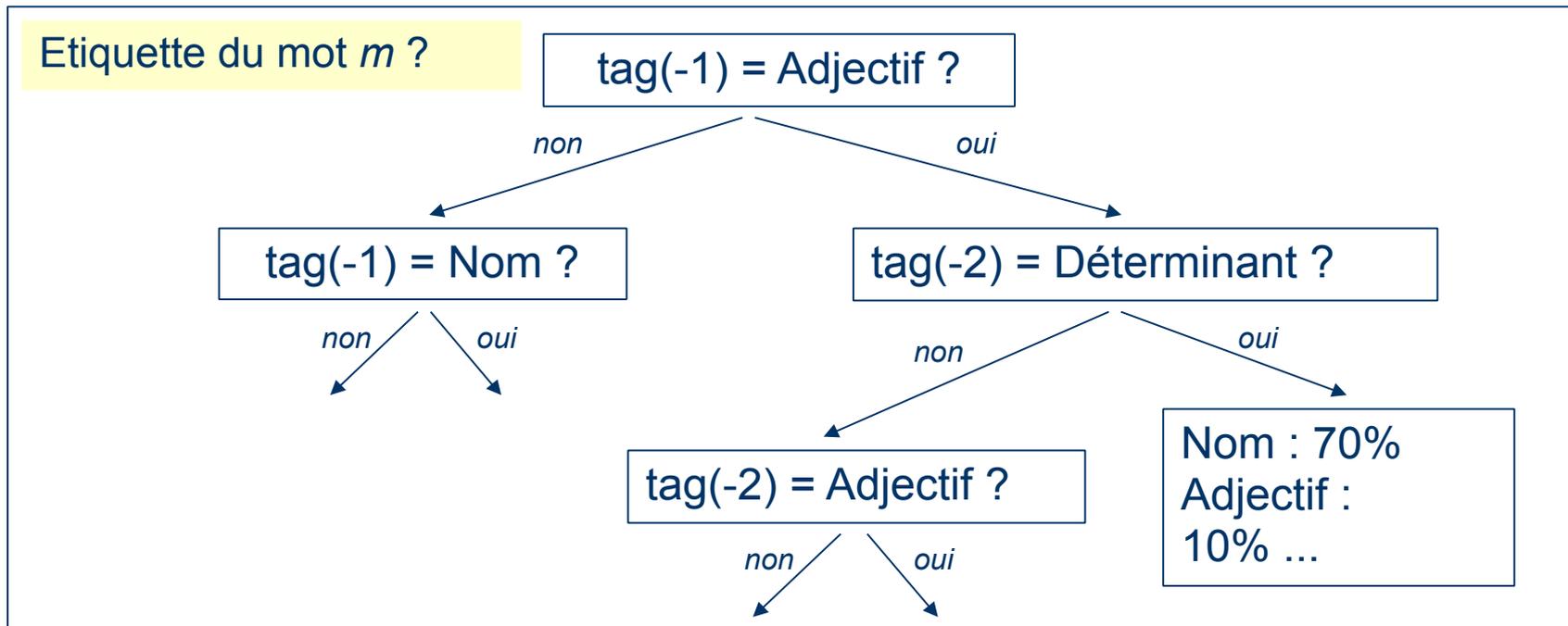
Mais/**COO** pour/**PREP**  
des/**DTN:p1**  
personnes/**SBC:p1**  
très/**ADV**  
spontanées/**ADJ**

...

# TreeTagger (1/2)

- **Généralités sur l'étiqueteur le TreeTagger [Schmid 1994]**
  - Estimation qu'un mot ait une étiquette grammaticale (Nom, Adjectif, etc.) en s'appuyant sur des arbres de décision binaires.
  - Les arbres sont construits récursivement à partir d'un ensemble de trigrammes connus (suite de 3 étiquettes consécutives constituant l'ensemble d'apprentissage).

# TreeTagger (2/2)



- $P(\text{tag}_m = \text{Nom} \mid \text{tag}(-2) = \text{déterminant}, \text{tag}(-1) = \text{Adjectif}) = 70\%$
- $P(\text{tag}_m = \text{Adjectif} \mid \text{tag}(-2) = \text{déterminant}, \text{tag}(-1) = \text{Adjectif}) = 10\%$

# Etiqueteur de Brill (2/3)

- **Généralités sur l'étiqueteur de Brill**

- Changement d'un tag par un autre suivant les tags des mots proches (tag des mots précédents ou suivants ou des deux mots précédents etc).

Exemple : ... *can/modal see/noun* ... -> ... *can/modal see/verb* ...

- Utilisation des contextes : changement d'un tag par un autre suivant les mots proches en présence (on ne prend pas en compte, comme précédemment, leur tag).

Exemple : ... *as/adverbe tall/adjective as/preposition* ...

-> ... *as/preposition tall/adjective as/preposition* ...

- Utilisation de certaines caractéristiques pour les mots inconnus (lettres majuscules pour les noms propres, suffixe des mots ...)

# Etiqueteur de Brill (1/3)

- **Généralités sur l'étiqueteur de Brill [Brill 1994]**
  - Le but est d'apprendre des règles d'étiquetage à partir d'un corpus annoté manuellement (« Wall Street Journal »).
  - A chaque étape d'apprentissage, des règles seront modifiées et le résultat de l'étiquetage avec ces nouvelles règles sera comparé avec le corpus représentant l'ensemble des annotations justes.
  - Tant qu'un nombre d'erreurs seuil dans l'étiquetage subsiste, le processus d'apprentissage continue.

# Etiqueteur de Brill (3/3)

- **Généralités sur l'étiqueteur de Brill**

- Lexique pas toujours adapté pour des textes spécialisés.

=> Améliorations de l'étiqueteur de Brill :

*Ajouter :*

- des règles lexicales et contextuelles propres au domaine
- ajout d'étiquettes spécifiques au domaine

# Evaluation de la qualité d'un étiquetage grammatical

- 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.**

# Evaluation de la qualité d'un étiquetage grammatical

- Mesures d'évaluation appliquées à une étiquette grammaticale (par exemple, *Nom*, *Adjectif*, etc.).

$$\textit{précision} (type\_étiquette) = \frac{\text{nombre d'étiquettes correctes appliquées} (type\_étiquette)}{\text{nombre d'étiquettes appliquées} (type\_étiquette)}$$

$$\textit{rappel} (type\_étiquette) = \frac{\text{nombre d'étiquettes correctes appliquées} (type\_étiquette)}{\text{nombre d'étiquettes correctes} (type\_étiquette)}$$

# Evaluation de la qualité d'un étiquetage grammatical : *Exemple*

Vous/PRV:pl faites/VCJ:pl **preuve/SBC:sg** de/PREP **mesure/SBC:sg** dans/PREP vos/DTN:pl **propos/SBC:pl** ,/, et/COO votre/DTN:sg **discours/SBC:sg** est/ECJ:sg toujours/ADV empreint/ADJ1PAR:sg de/PREP **réserve/SBC:sg** ./.

Vous/PRV:pl n'/ADV êtes/ECJ:pl certainement/ADV pas/ADV **indifférent/SBC:sg** ,/, mais/COO peu/ADV **expansif/SBC:pl** ./.

Votre/DTN:sg **approche/SBC:sg** plutôt/ADV **formaliste/SBC** peut/VCJ:sg amener/VNCFF vos/DTN:pl **interlocuteurs/SBC:pl** à/PREP penser/VNCFF que/SUB vous/PRV:pl portez/VCJ:pl une/DTN:sg grande/ADJ:sg **attention/SBC:sg** aux/DTC:pl **conventions/SBC:pl** ou/COO aux/DTC:pl **usages/SBC:pl** ./.

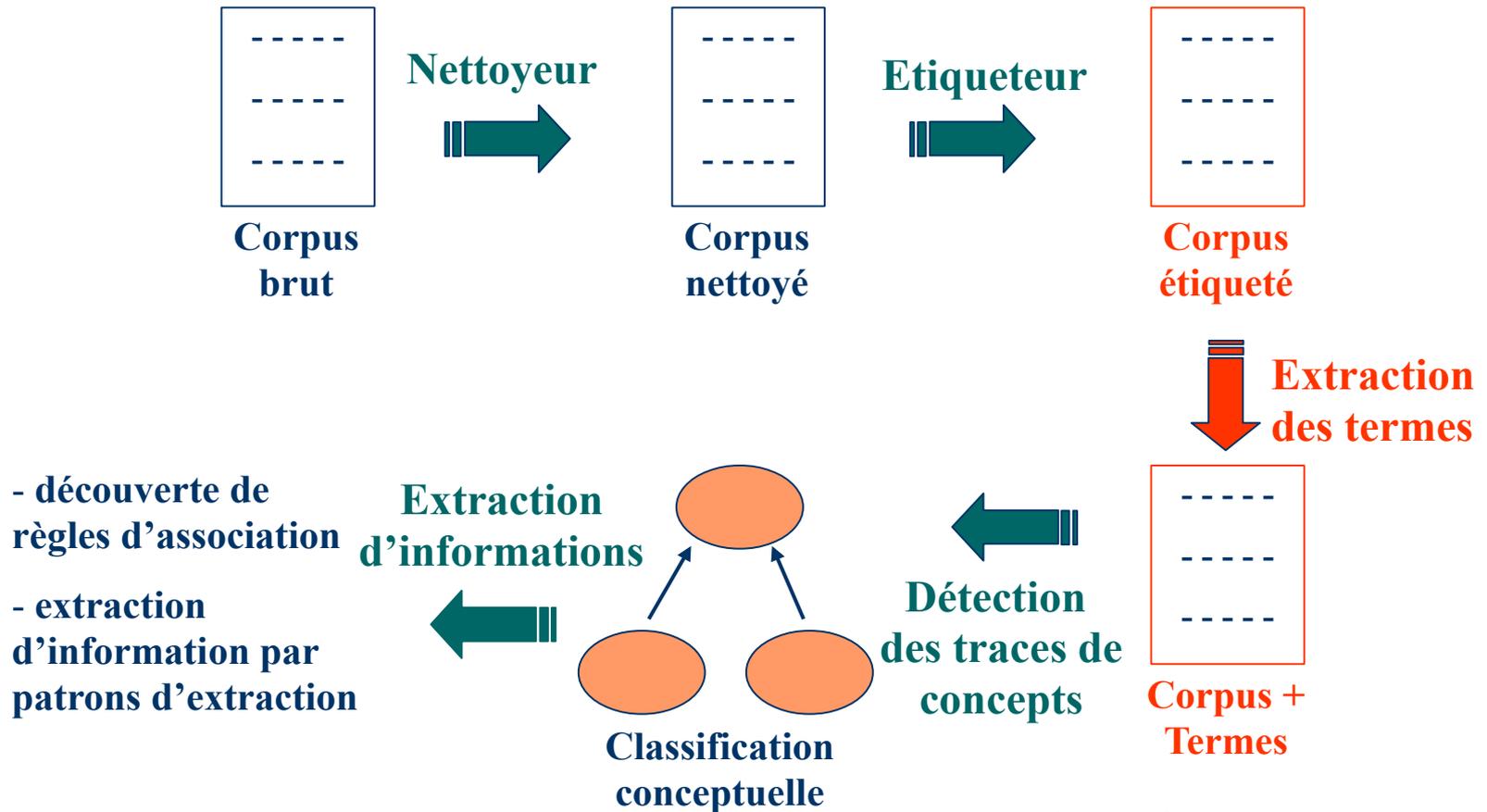
Votre/DTN:sg **comportement/SBC:sg** peut/VCJ:sg ,/, par/PREP contre/PREP ,/, paraître/VNCFF assez/ADV fermé/ADJ2PAR:sg à/PREP ceux/PRO:pl qui/REL ont/ACJ:pl coutume/ADJ:sg de/PREP réagir/VNCFF spontanément/ADV ./.

Votre/DTN:sg **approche/SBC:sg** sérieuse/ADJ:sg peut/VCJ:sg amener/VNCFF vos/DTN:pl **interlocuteurs/SBC:pl** à/PREP penser/VNCFF que/SUB vous/PRV:pl considérez/VCJ:pl le/DTN:sg **temps/SBC:sg** comme/SUB un/DTN:sg...

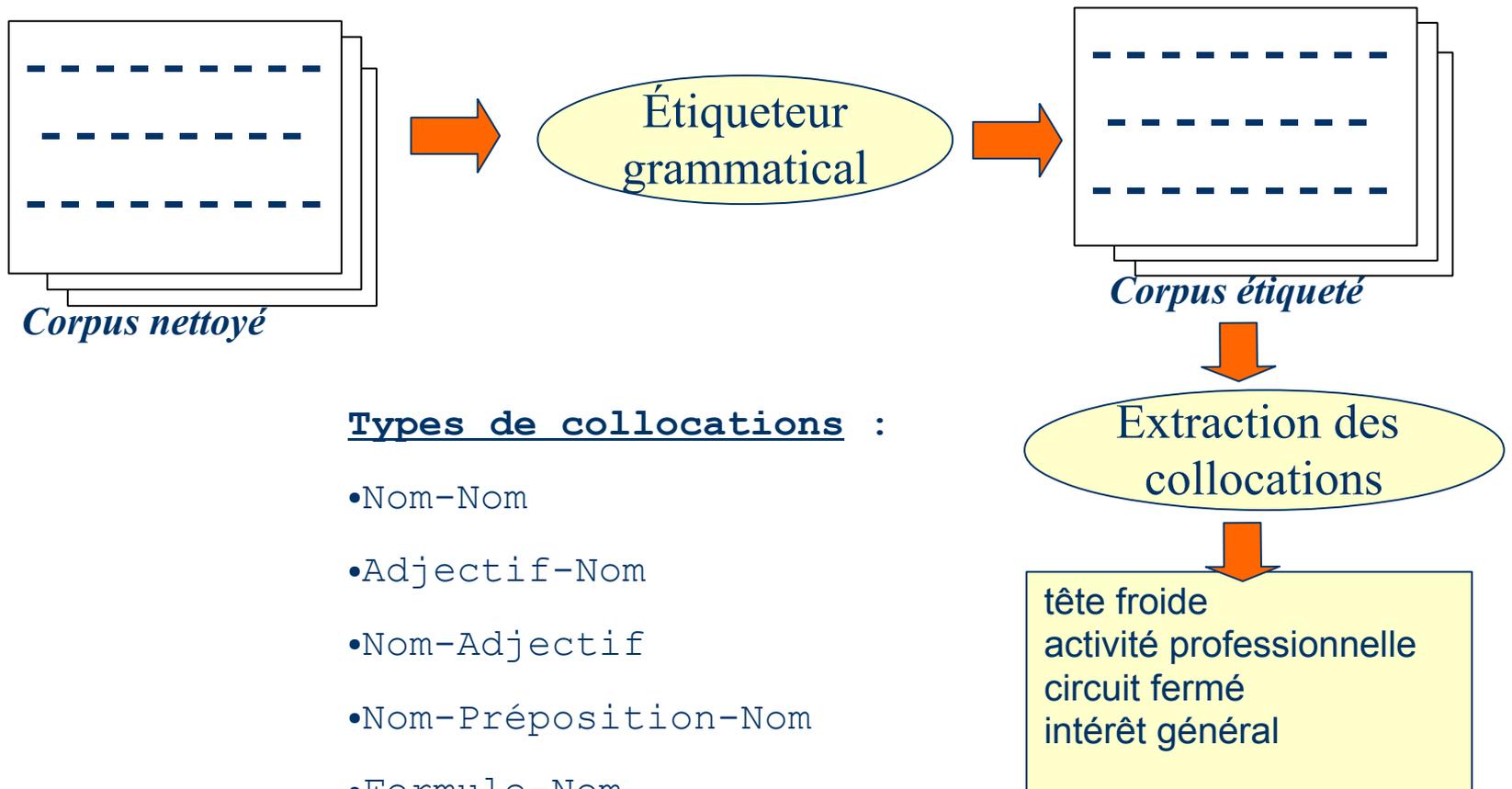
# Evaluation de la qualité d'un étiquetage grammatical : *Exemple*

- Calculer le rappel et la précision des noms (étiquettes "SBC" sur l'exemple).
- Comment obtenir un rappel de 100 % ? Dans ce cas, quelle est la précision ?
- Comment obtenir une précision de 100% ? Dans ce cas, quel est le rappel ?

# Processus de fouille de textes



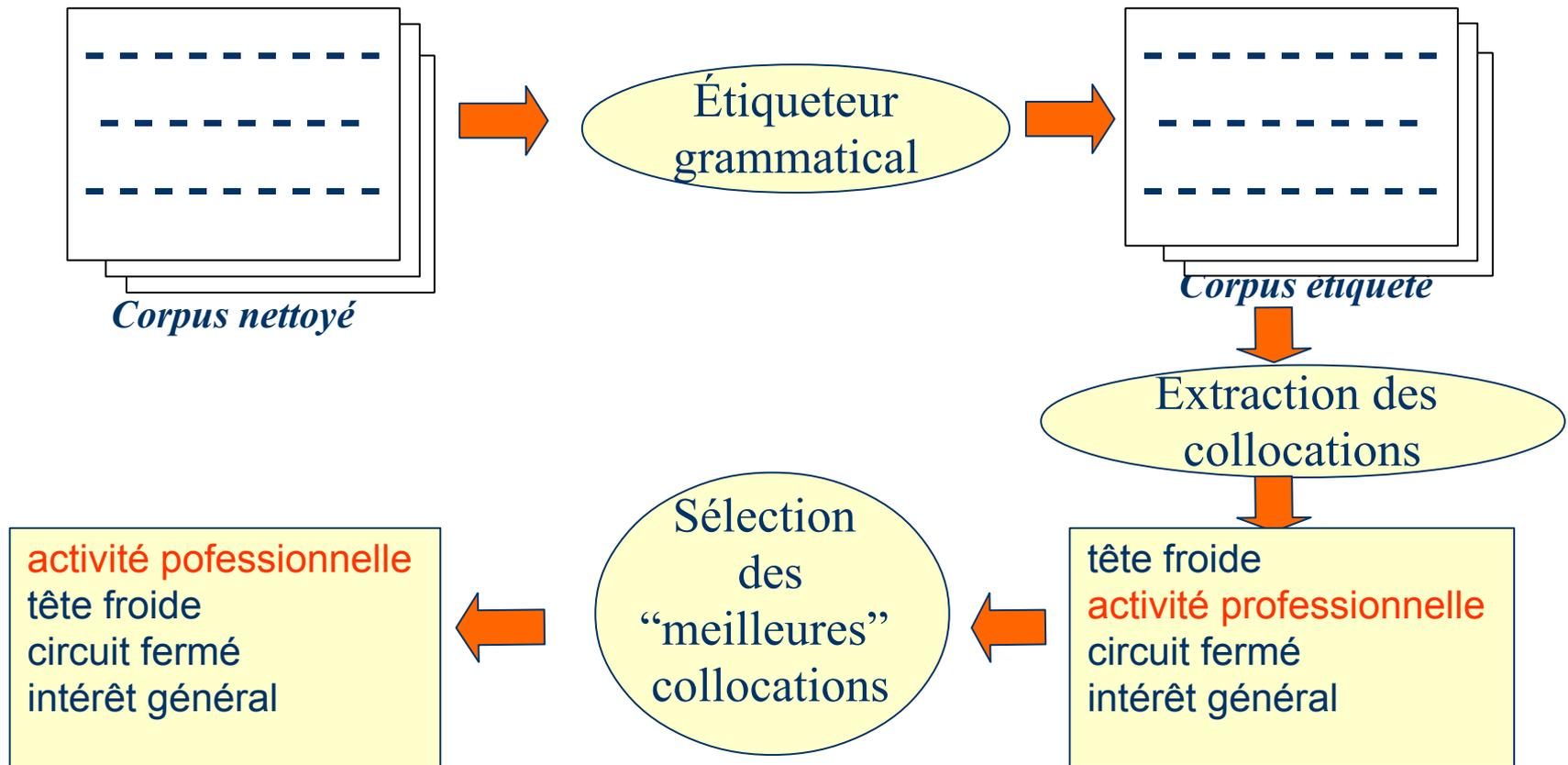
# Etape 3 : Extraction des termes



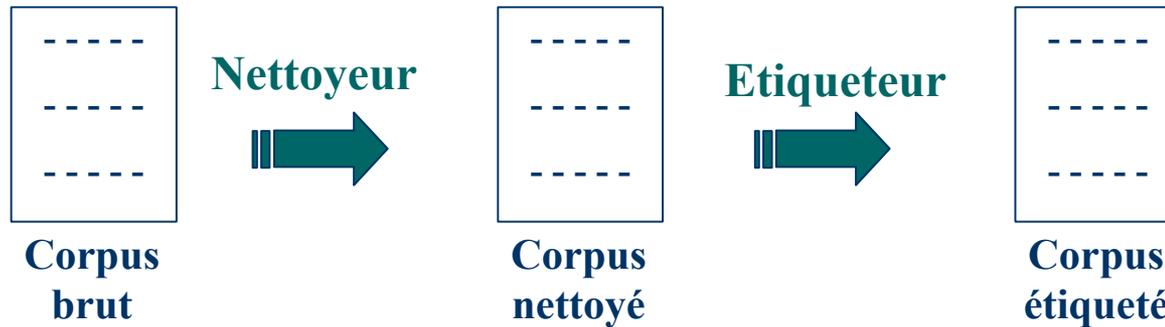
## Types de collocations :

- Nom-Nom
- Adjectif-Nom
- Nom-Adjectif
- Nom-Préposition-Nom
- Formule-Nom ...

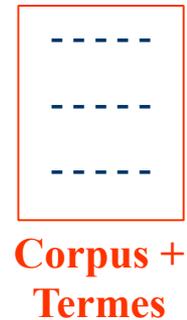
# Etape 3 : Extraction des termes



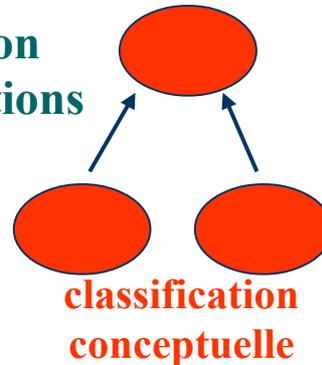
# Processus de fouille de textes



Extraction des termes



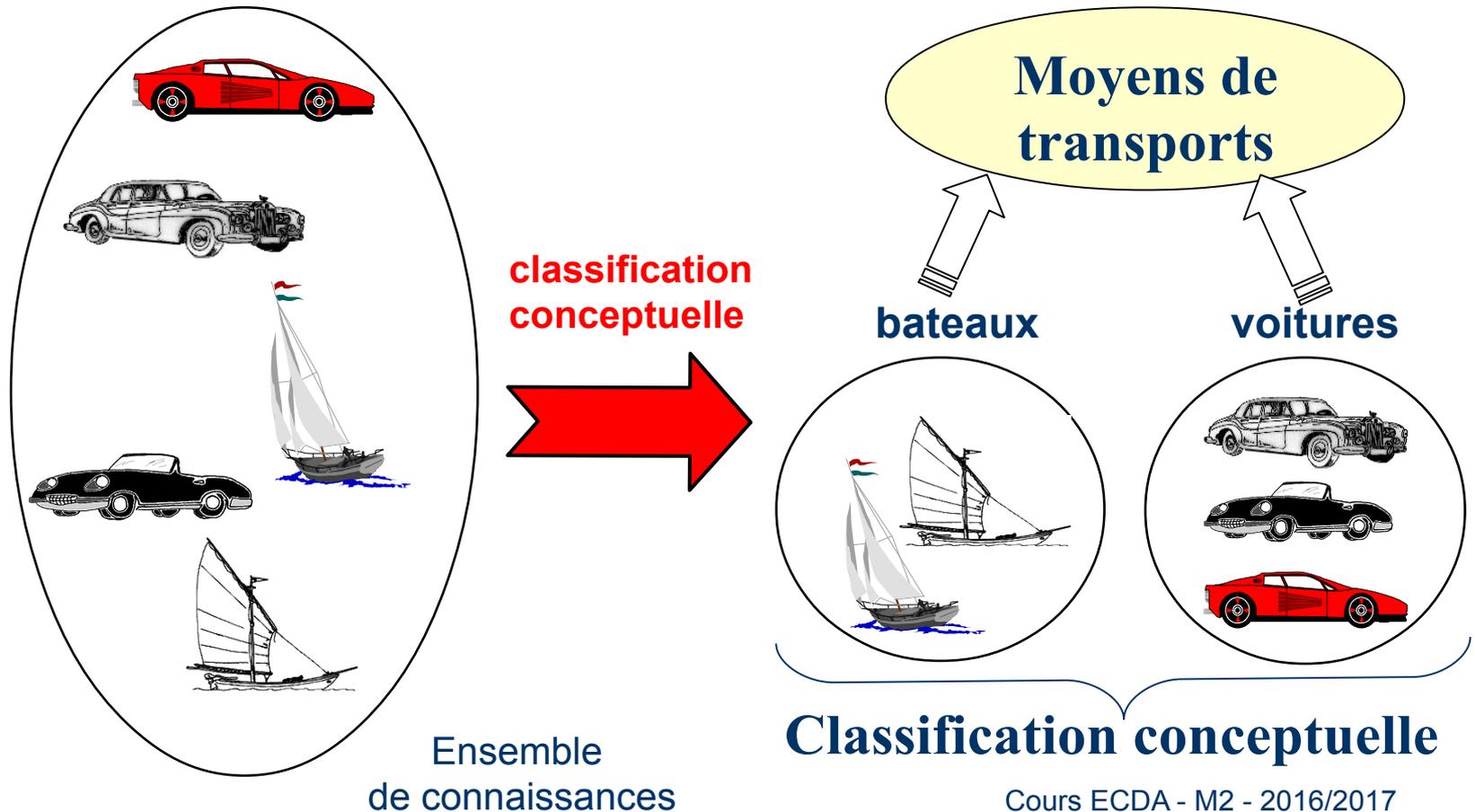
Détection des traces de concepts



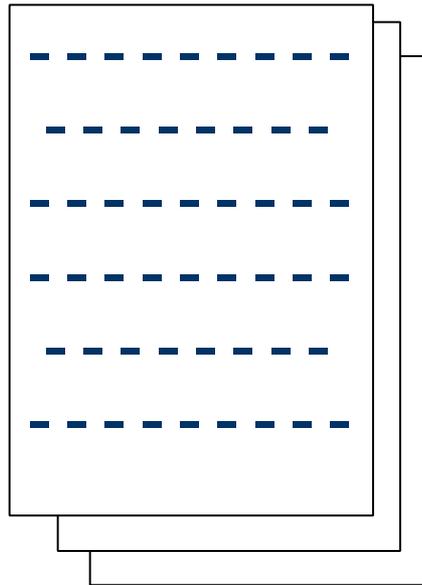
Extraction d'informations

- découverte de règles d'association
- extraction d'information par patrons d'extraction

# Classification conceptuelle



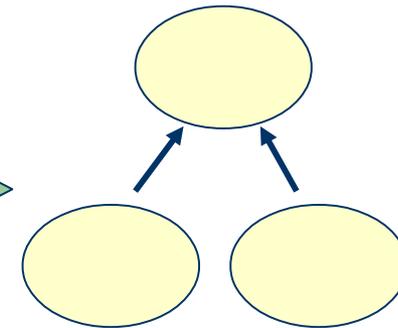
# Etape 4 : Détection des traces de concepts



*Corpus avec prise en compte de la terminologie*



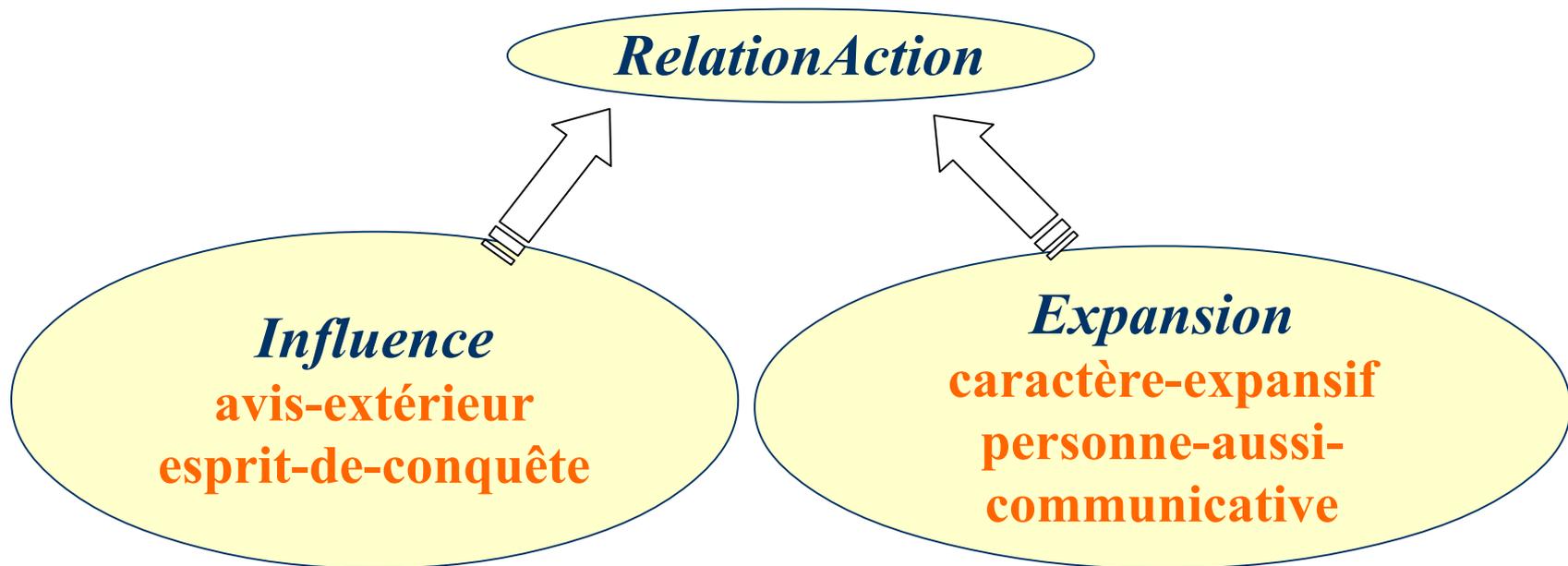
*LSA, Asium, etc.*



*Classification conceptuelle*

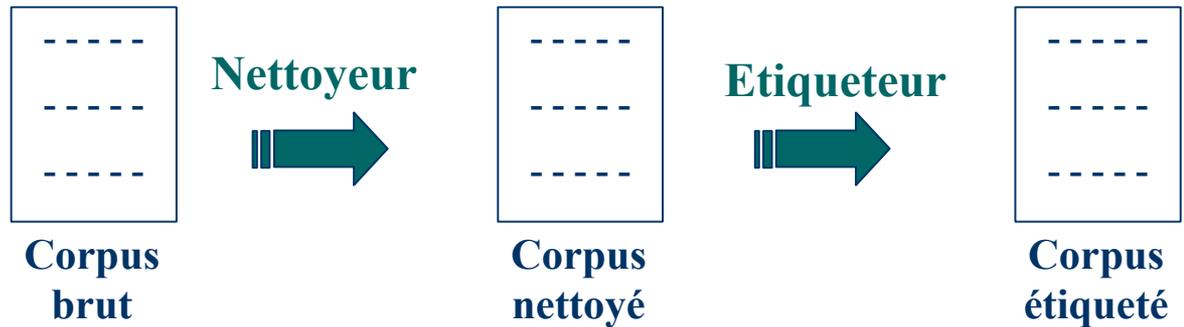
# Classification conceptuelle

- Exemple de classification spécialisée (*construite à partir d'un corpus des Ressources Humaines*)

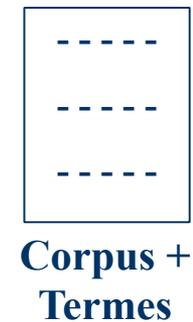


- Classification généraliste : **WordNet**

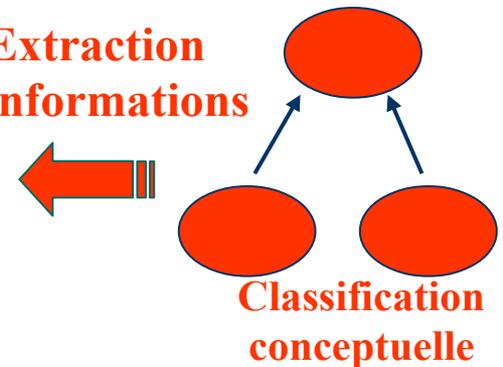
# Processus de Fouille de textes



Extraction des termes



Détection des traces de concepts



Extraction d'informations

- découverte de règles d'association
- extraction d'information par patrons d'extraction

# Etape 5 : Extraction d'informations

- Extraction d'informations par patrons d'extraction

*Exemple:*

..MSN2 encode a **zinc-finger transcriptional activator** , ...

..MSN4 encode a **DNA-binding component of the stress responsive system** , ...

**2 patrons d'extraction sont nécessaires** pour rechercher la spécificité des protéines codées par les gènes de régulation de transcription :

- **MSN2 encode** *SpécificitéFacteur*
- **MSN4 encode** *SpécificitéFacteur*

# Etape 5 : Extraction d'informations

- Extraction d'informations par patrons d'extraction

*Exemple:*

...MSN2 encode a **zinc-finger transcriptional activator** , ...

...MSN4 encode a **DNA-binding component of the stress responsive system** , ...

**1 seul patron d'extraction suffit** pour rechercher la spécificité des protéines codées par les gènes de régulation de transcription avec la **connaissance sémantique**.

- **\$TranscriptionActivator** encode **SpécificitéFacteur**

## Etape 5 : Extraction d'informations

- Extraction de règles d'associations

bending-influence (nom-verbe)

*Bendng*

DNA-duplex

*DNAconformatn*

transcription-factor

*Regulfactor*

gal4-binding

*Regulfactor*

interaction-with-TFIIB

*Transcriptn*

**Bendng, DNAconformatn, Regulfactor → Transcriptn**

# Bilan

