

# Une chaîne globale de fouille de textes

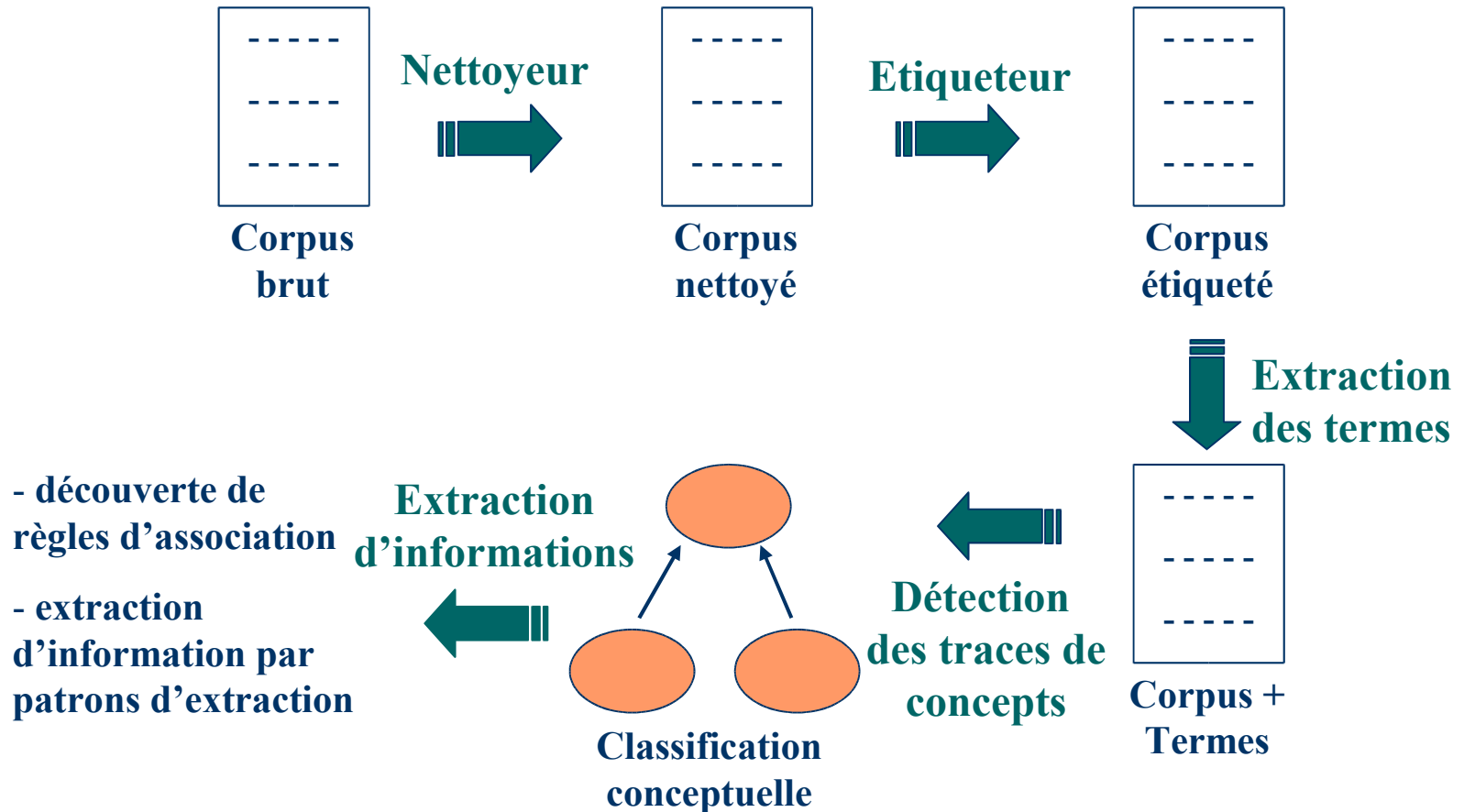
Mathieu Roche

Cours ECD

---

2007/2008

# 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

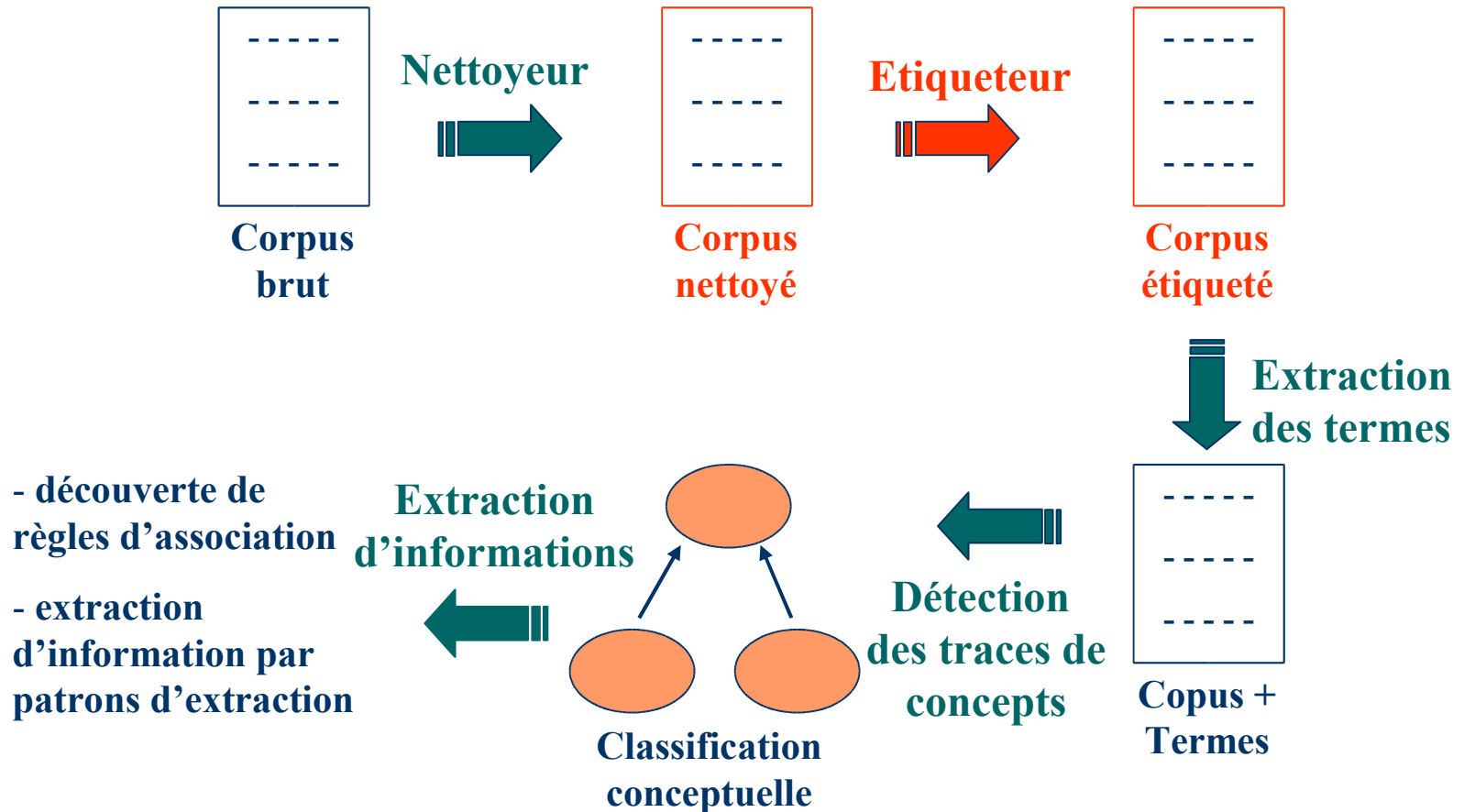
**CORPUS FOUILLE DE DONNEES** : Remplacer ([lettres+année], [numéro], etc.) par « a paper » ou « papers » si ces références sont précédées de la préposition « in », sinon on supprime ces références.

- 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 ...

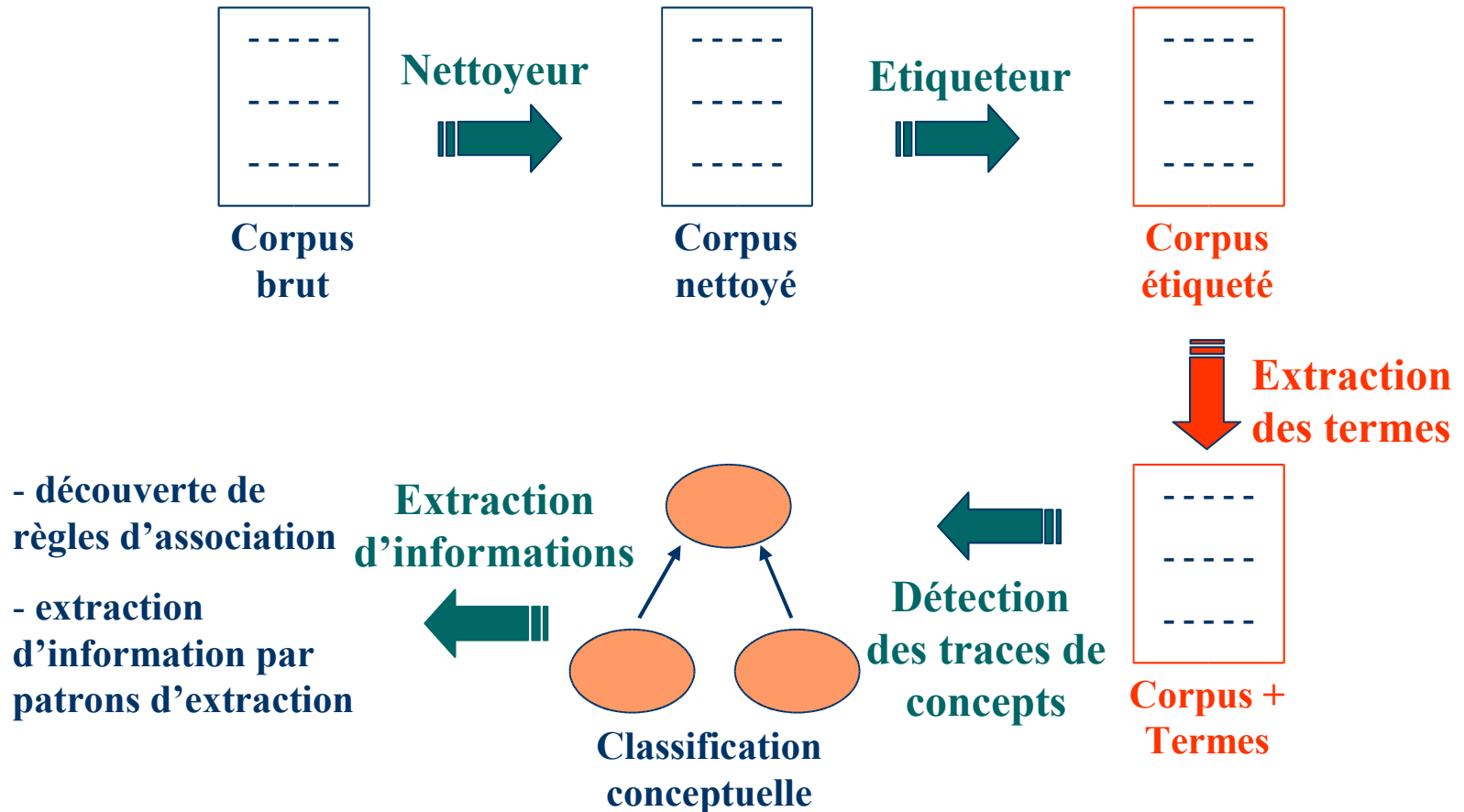


**Étiqueteur  
de Brill**

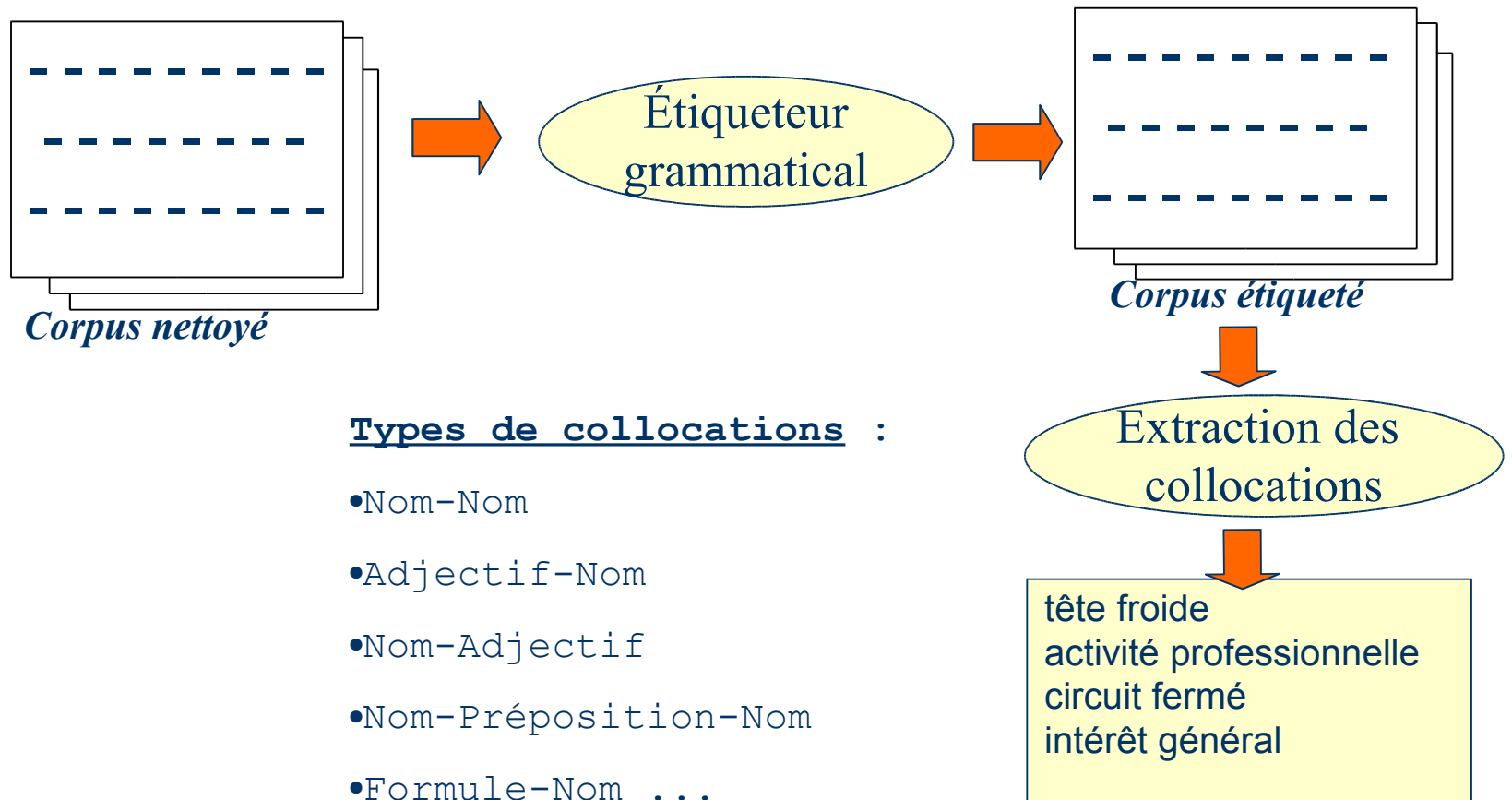
Mais/**COO** pour/**PREP**  
des/**DTN:p1**  
personnes/**SBC:p1**  
très/**ADV**  
spontanées/**ADJ**

...

# Processus de fouille de textes



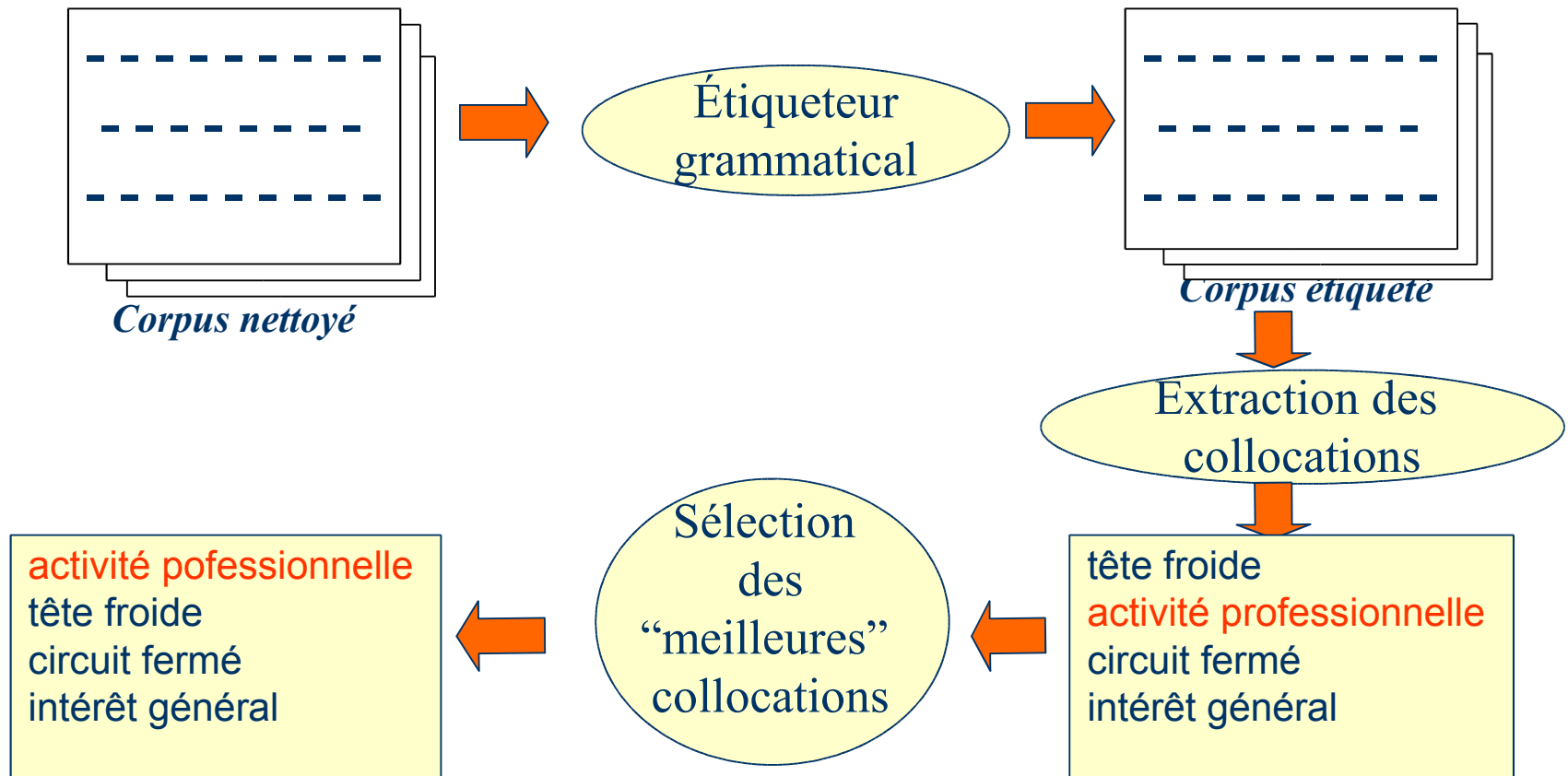
# Etape 3 : Extraction des termes



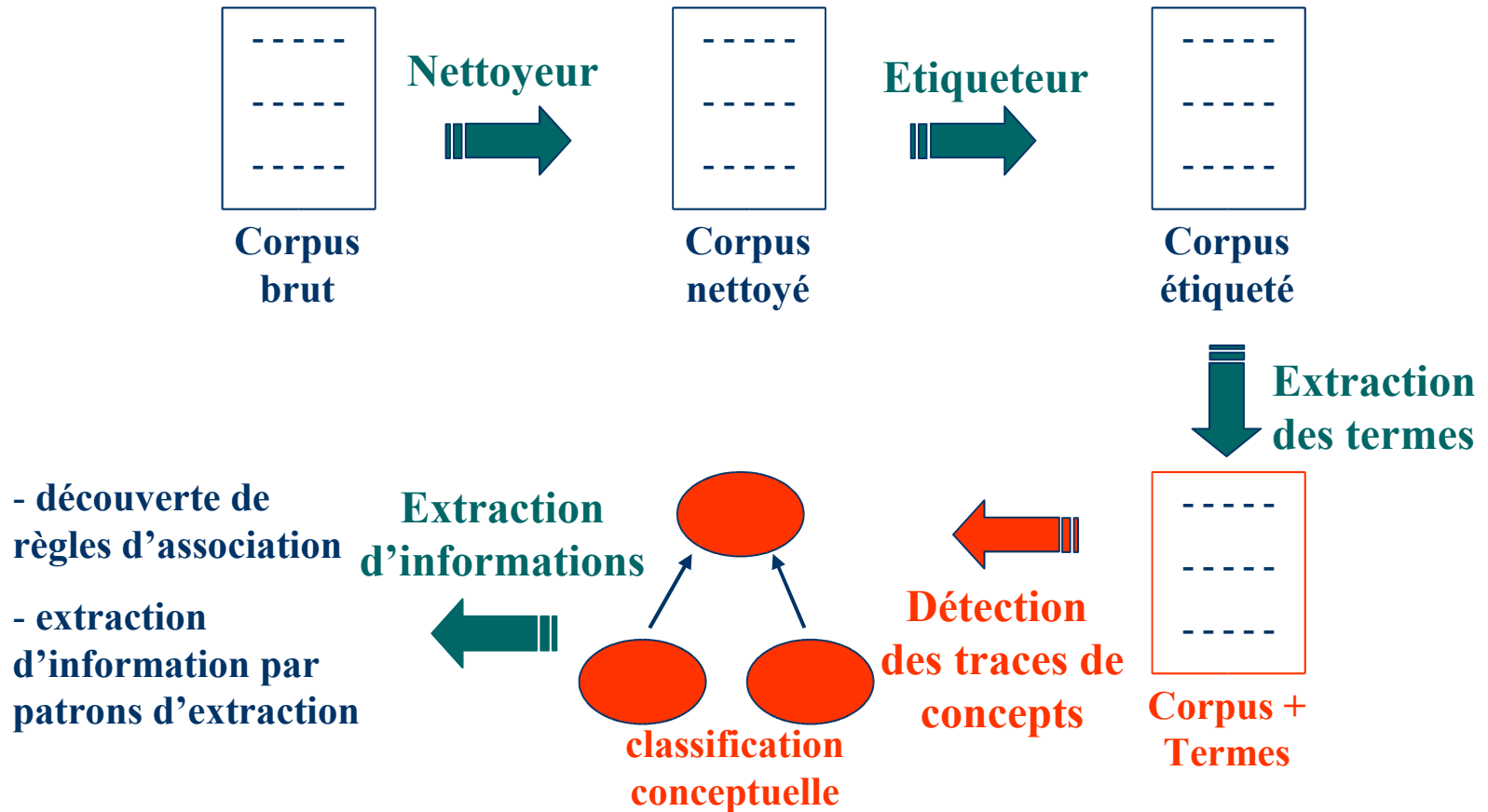
## Types de collocations :

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

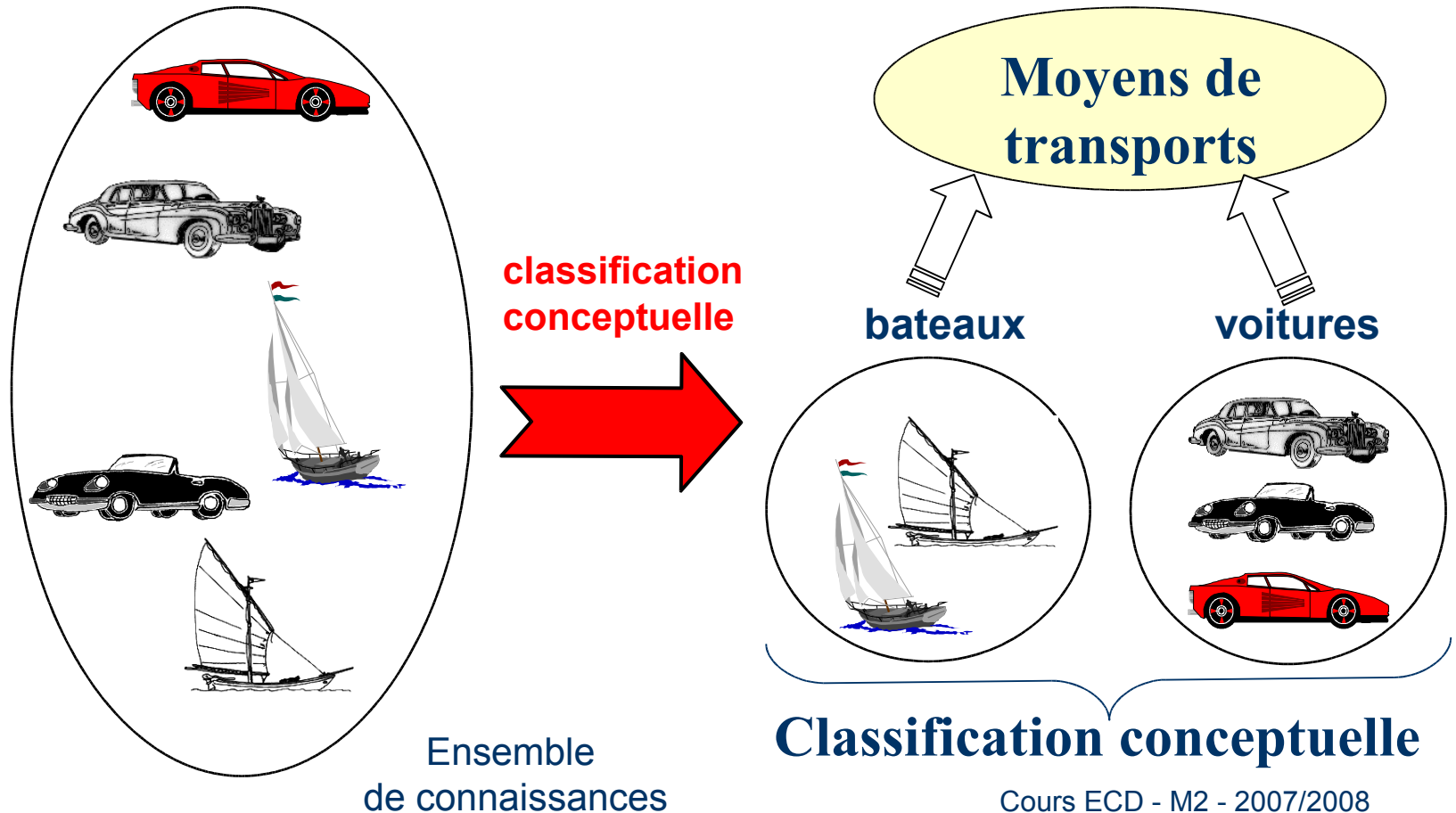
# Etape 3 : Extraction des termes



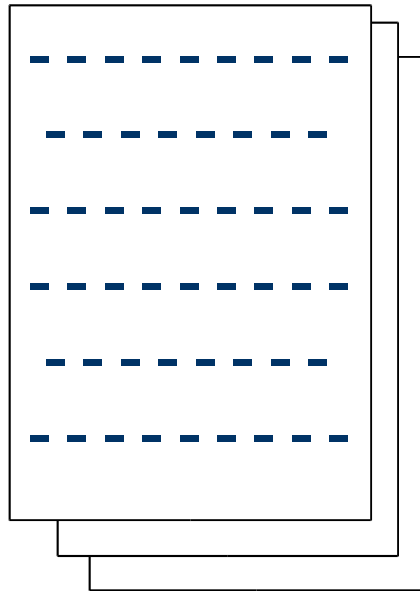
# Processus de fouille de textes



# Classification conceptuelle



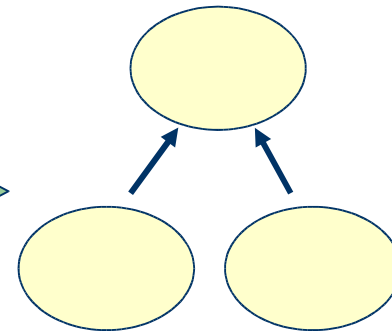
# Etape 4 : Détection des traces de concepts



*Corpus avec prise en compte de la terminologie*



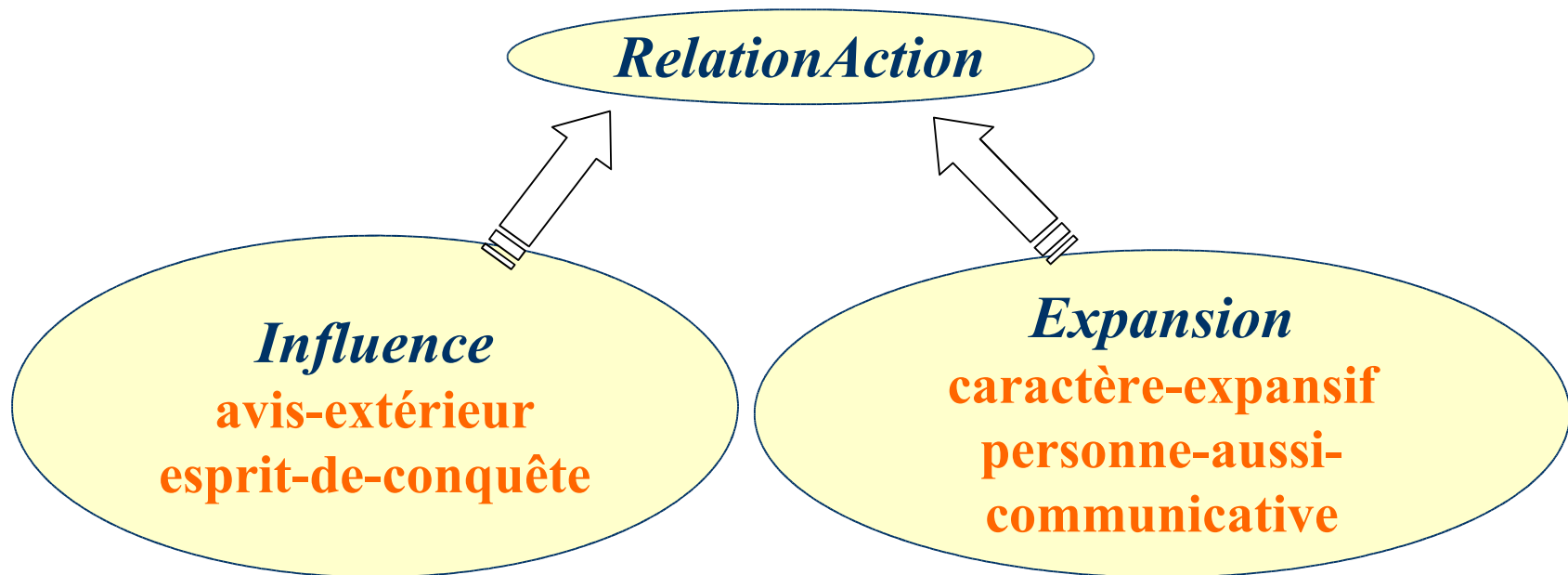
*LSA, Asium, etc.*



*Classification conceptuelle*

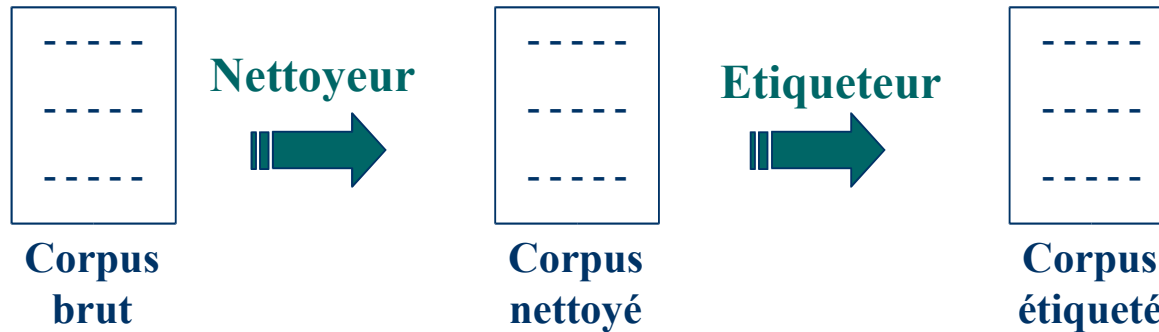
# Classification conceptuelle

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

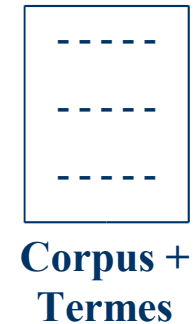


- Classification généraliste : **WordNet**

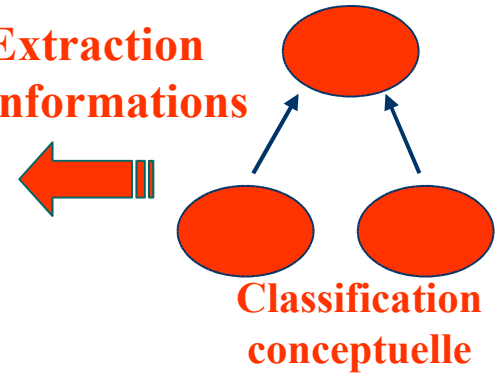
# Processus de Fouille de textes



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

