

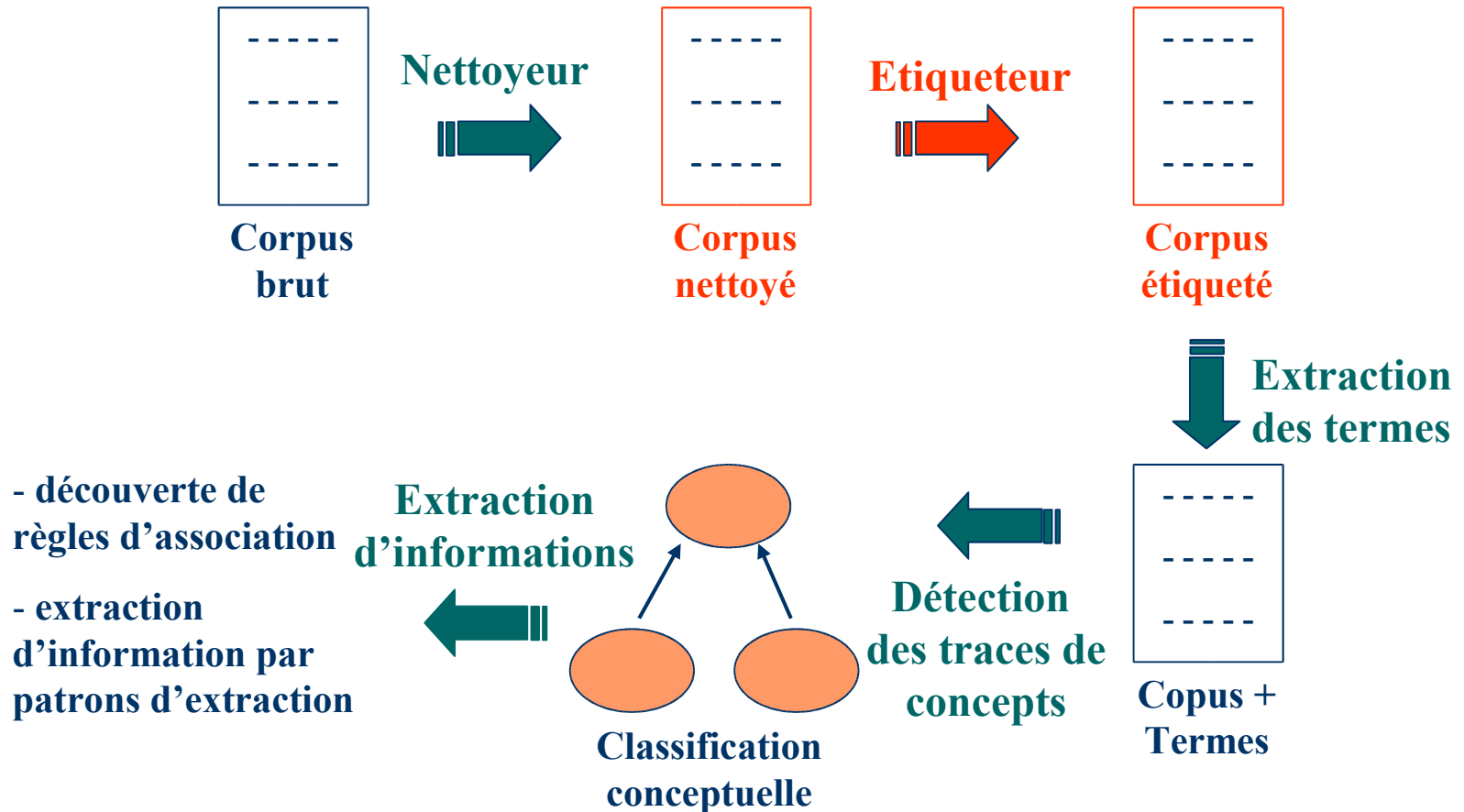
# Etiquetage

**Mathieu Roche**

**Cours ECD**

**2008/2009**

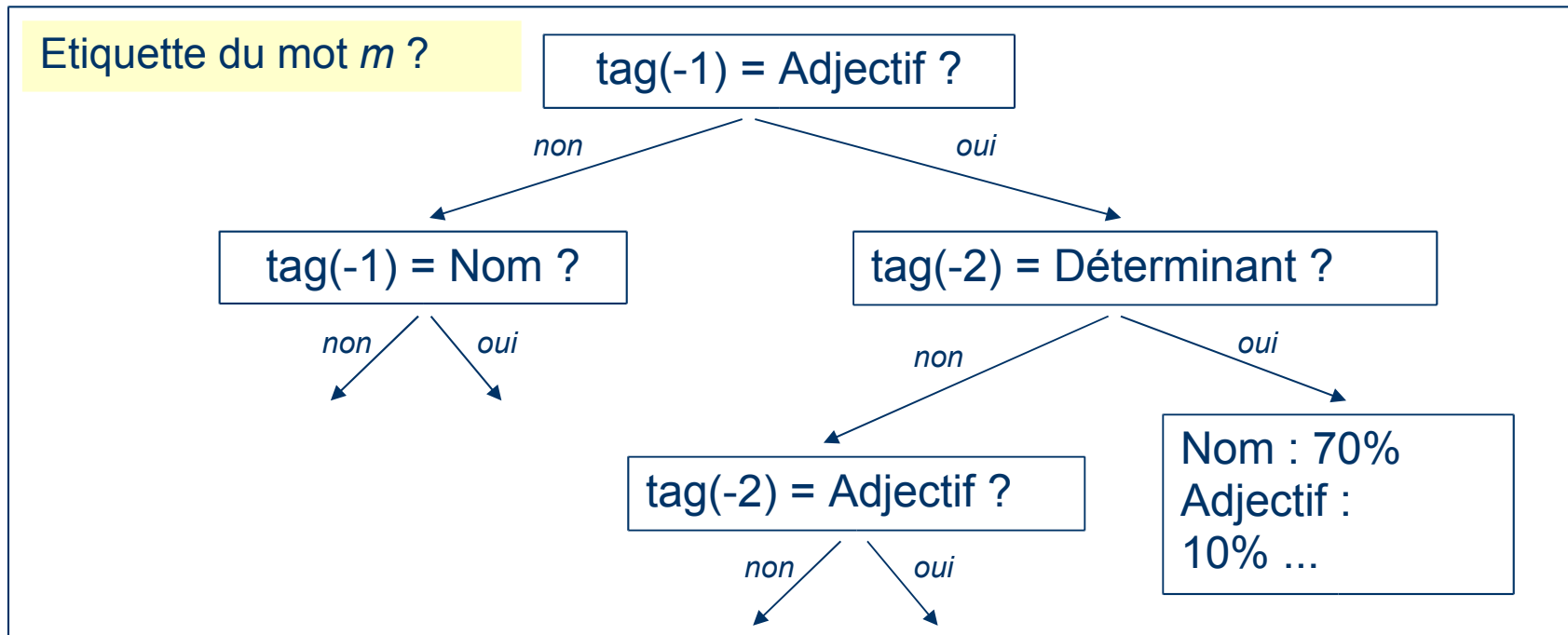
# Processus de fouille de textes



# TreeTagger <sup>(1/2)</sup>

- **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 tigrammes 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 (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 (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 (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.).

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

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

# 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 ?

# Evaluation de la qualité d'un étiquetage grammatical

- Méthode pour combiner Rappel et Précision : le Fscore.

$$Fscore = \frac{(\beta^2 + 1) \times Précision \times Rappel}{(\beta^2 \times Précision) + Rappel}$$

- Avec  $\beta=1$ , calculer le Fscore pour chacun des trois cas précédents. Conclure.

# Exemple correctement étiqueté

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/ADJ:sg ,/, mais/COO peu/ADV expansif/ADJ:sg ./.

Votre/DTN:sg **approche/SBC:sg** plutôt/ADV formaliste/ADJ:sg 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/SBC: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...