

GTRC Modéliser le changement: les voies du français

Rencontre de mi-parcours, Ottawa, 13-17 août 2007

Perspectives sur les outils informatiques:
TreeTagger

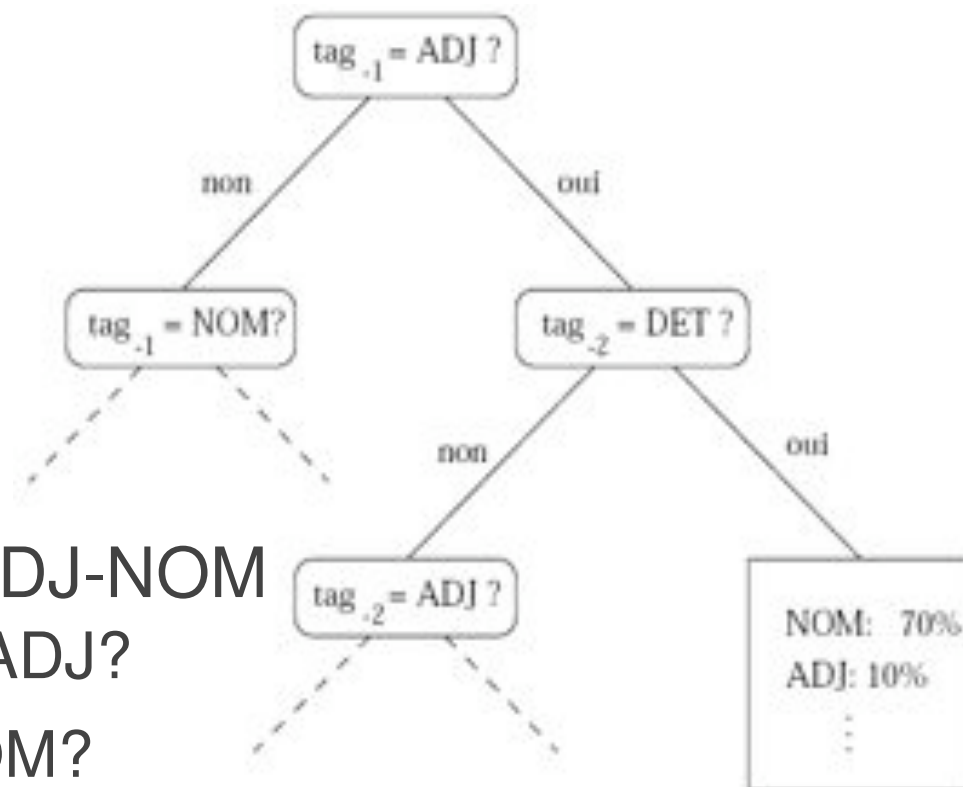
Achim Stein (Stuttgart)

TreeTagger

- ▶ "TreeTagger": un étiqueteur probabiliste
 - ▶ Algorithme: Helmut Schmid (IMS, Stuttgart)
 - ▶ Particularités (comparé aux modèles Hidden Markov traditionnels):
 - ▶ Calcul des probabilités de transition moyennant un arbre de décision
 - ▶ Contexte variable
 - ▶ Analyse d'affixe
- ▶ Entraînement
 - ▶ Lexique et corpus d'entraînement annoté
- ▶ Résultat: lexique paramétrisé (fichier de paramètres)
 - ▶ Anglais, allemand, autres langues
 - ▶ Français
 - ▶ Ancien français

Fonctionnement du TreeTagger

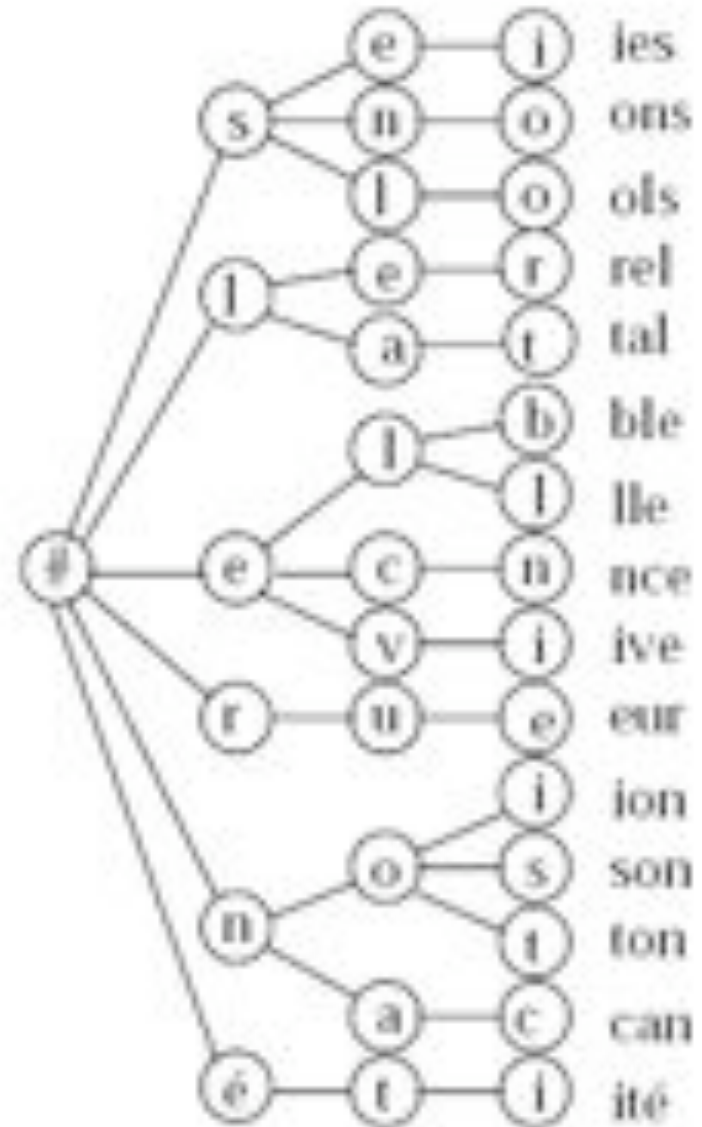
- ▶ Arbres de décisions
 - ▶ décisions binaires
 - ▶ p.ex. trigramme DET-ADJ-NOM
 - ▶ mot précédent (tag_{-1}): ADJ?
 - ▶ premier mot (tag_{-2}): NOM?
 - ▶ la feuille contient les probabilités
- ▶ Avantages:
 - ▶ contextes variables
 - ▶ spécification de contextes négatifs:
 - ▶ $tag_{-1}=ADJ$ & $tag_{-2}=ADJ$ & $tag_{-3}=DET$



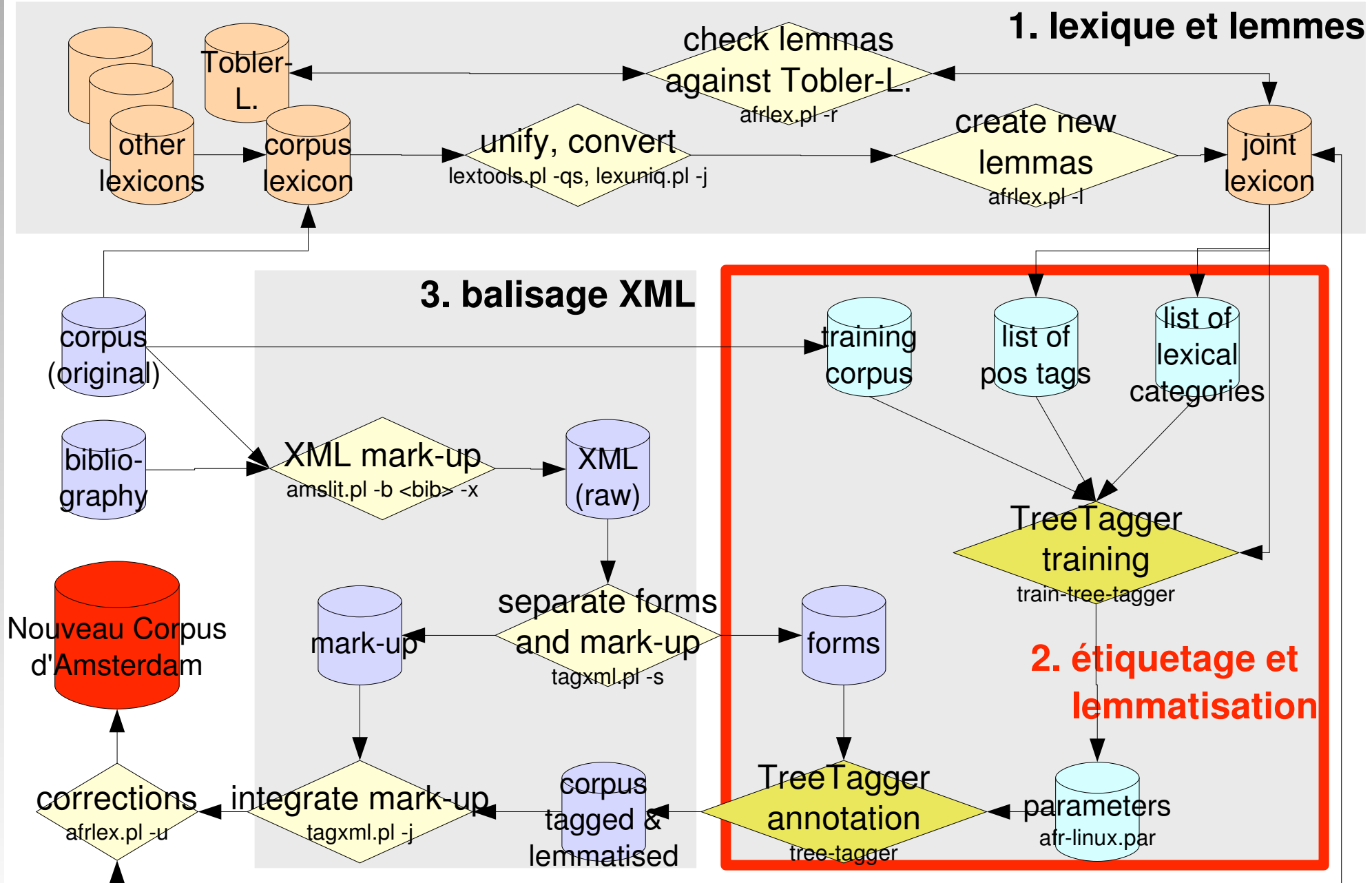
Traitement des formes inconnues

- ▶ Recours aux arbres "d'affixes" si la forme ne figure pas dans le lexique
- ▶ les probabilités (étiquette position et suffixe) sont alors multipliées et normalisées:

$$p_a(t | w) := p_p(t | [w]_p) p_s(t | [w]_s) n(w)$$



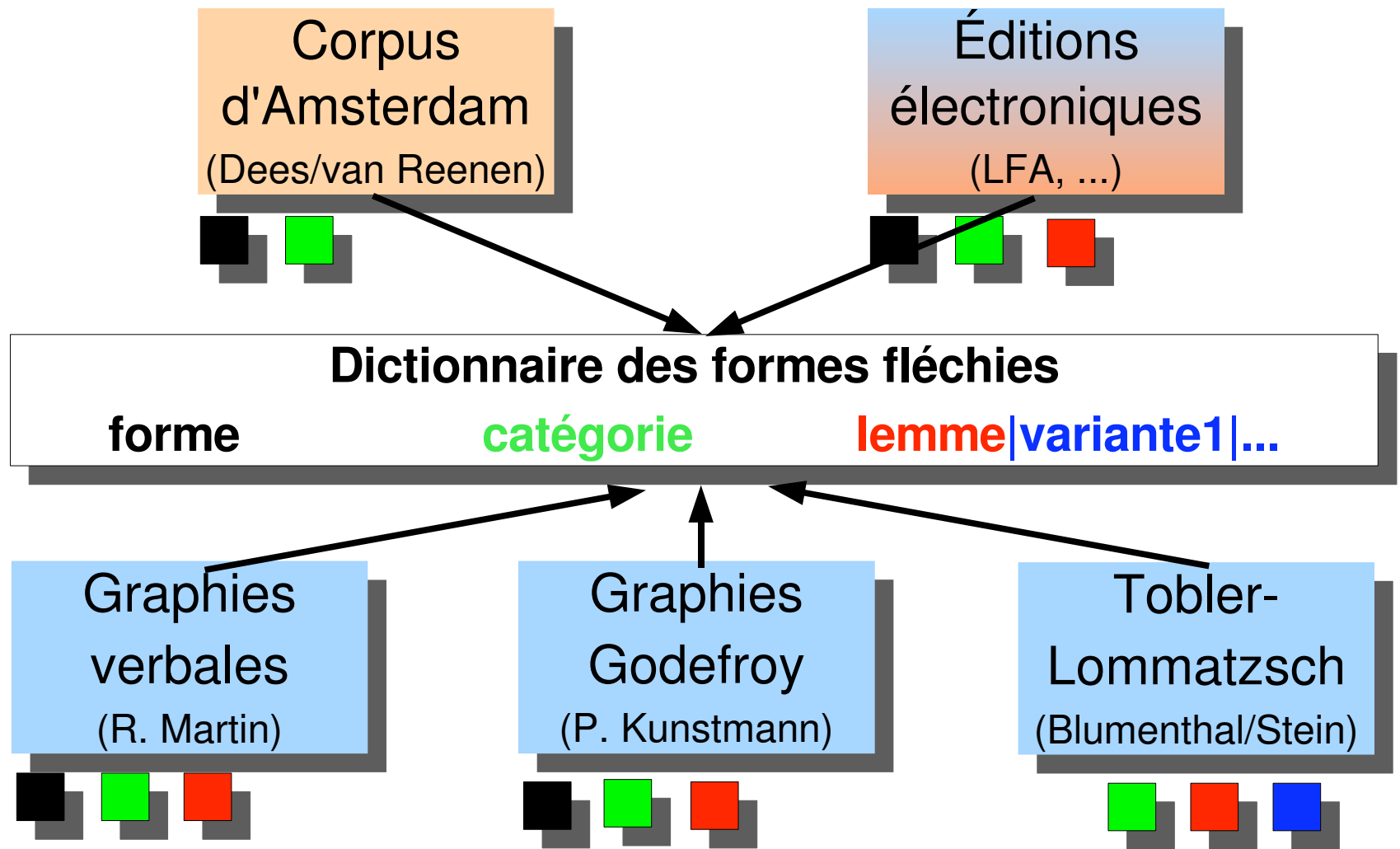
Annotation du Corpus d'Amsterdam



TreeTagger: utilisation

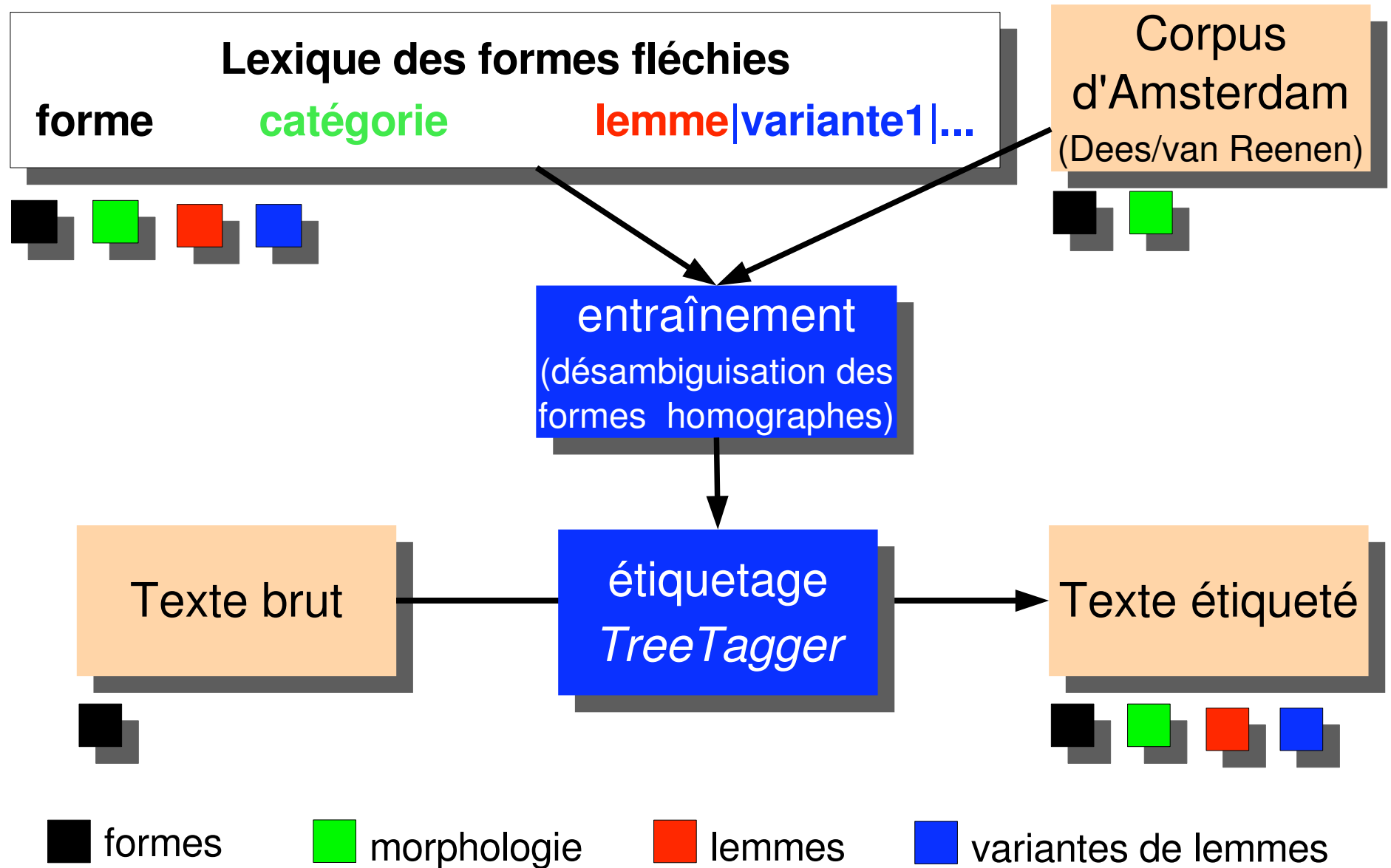
- ▶ Préalables de l'utilisation du TreeTagger
 - ▶ Jeu d'étiquettes
 - ▶ lexique des formes fléchies
 - ▶ avec étiquette (et lemme)
 - ▶ Corpus d'entraînement
 - ▶ annoté manuellement
- ▶ Utilisation des logiciels
 - ▶ Entraînement: `train-tree-tagger`
 - ▶ produit le fichier des paramètres (lexique intégré)
 - ▶ annotation: `tree-tagger`
 - ▶ produit le texte annoté étiqueté (et lemmatisé)

Ressources *textuelles* et *lexicales*



■ formes ■ morphologie ■ lemmes ■ variantes de lemmes

Étiquetage morphologique



TreeTagger: résultats (ancien français)

- ▶ Entraînement sur 2.685.146 mots
- ▶ Évaluation sur 500.000 mots
- ▶ Jeux d'étiquettes réduit à la partie du discours
 - ▶ Précision: 92,7%
 - ▶ Taux de lemmatisation: 97,8% des mots

| catégorie | types | tokens |
|--------------|--------------------|----------------------|
| ADJ | 2129/2774= 76.75% | 24009/24795= 96.83% |
| ADV | 1012/1285= 78.75% | 36514/36923= 98.89% |
| CON | 136/ 136=100.00% | 26497/26497=100.00% |
| DET | 264/ 264=100.00% | 49427/49427=100.00% |
| GDF | 8/ 8=100.00% | 16/ 16=100.00% |
| INT | 16/ 28= 57.14% | 190/ 251= 75.70% |
| NOM | 6915/9483= 72.92% | 76113/80050= 95.08% |
| NPR | 657/2185= 30.07% | 6206/ 8800= 70.52% |
| PON | 11/ 11=100.00% | 14212/14212=100.00% |
| PRE | 177/ 177=100.00% | 42914/42914=100.00% |
| PREDET | 29/ 29=100.00% | 6838/ 6838=100.00% |
| PRO | 438/ 438=100.00% | 79774/79774=100.00% |
| PROCON | 18/ 18=100.00% | 21832/21832=100.00% |
| VER | 14336/7282= 82.95% | 104439/07652= 97.02% |
| total | 76.60% | 97.80% |