# Oublier ce qu'on sait, pour mieux apprendre ce qu'on ne sait pas

Nicolas Pécheux

nicolas.pecheux@limsi.fr

en collaboration avec Alexandre Allauzen, Thomas Lavergne, Guillaume Wisniewski et François Yvon

14 avril 2015



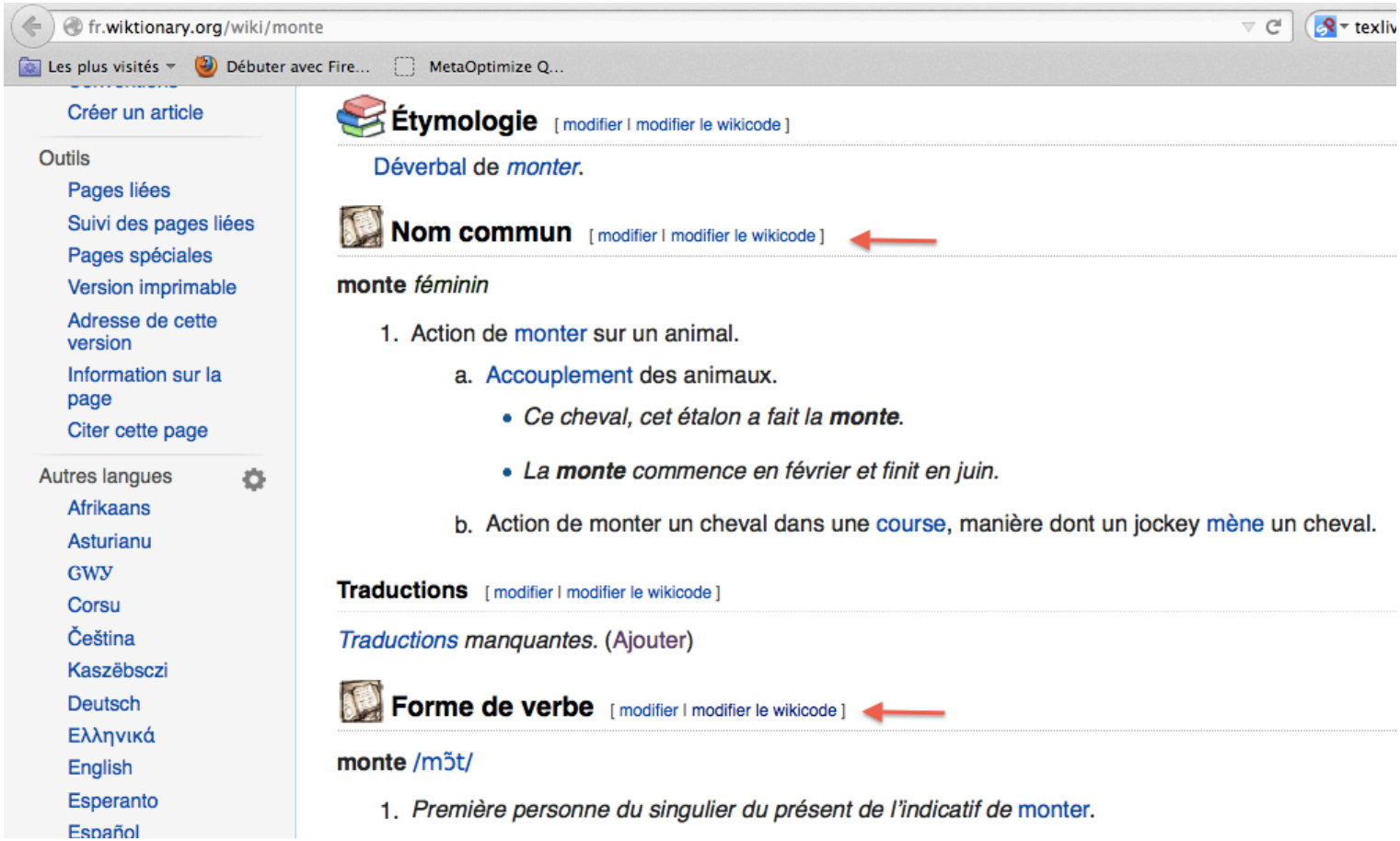
LIMSI-CNRS — Groupe TLP

# Contexte

- On considère une tâche d'analyse morpho-syntaxique (problème d'étiquetage de séquences)

# Motivation I

- On dispose de connaissances linguistiques, e.g. WIKTIONNAIRE



# Motivation II

- Le nombre d'étiquettes possibles à chaque position est trop grand

Un marché pour

cs=NOUN, cas=✗, num=sg, gen=masc, pers=✗, tmp=✗, mode=✗

cs=VERB, cas=✗, num=sg, gen=masc, pers=✗, tmp=✗, mode=pp

cs=VERB, cas=✗, num=pl, gen=masc, pers=✗, tmp=✗, mode=pp

cs=VERB, cas=✗, num=sg, gen=fem, pers=✗, tmp=✗, mode=pp

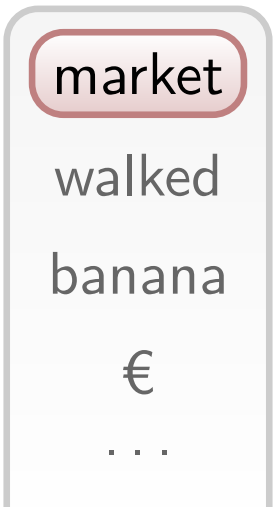
...

$> 10^3$

# Motivation (cachée) III

- Traduction automatique vue comme un problème d'étiquetage

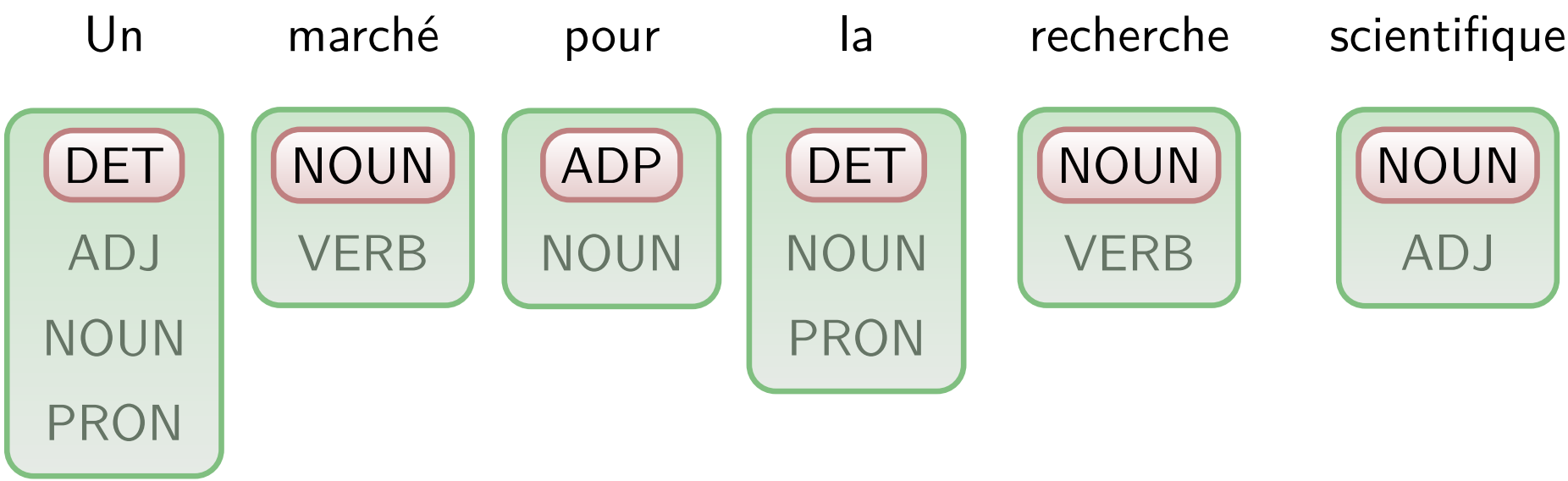
Un marché pour




$> 10^6$

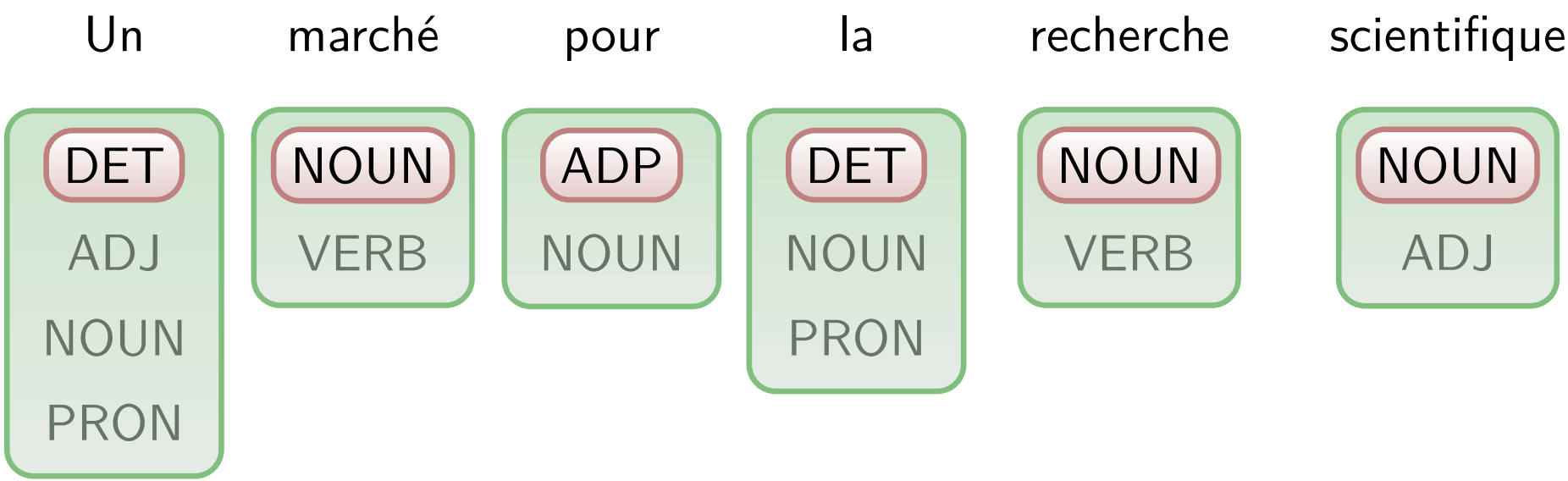
# Idée

- Utiliser des contraintes pour réduire l'espace des possibles
- Par exemple ici un dictionnaire { mot $\rightarrow$ étiquette }



# Idée

- Utiliser des contraintes pour réduire l'espace des possibles
- Par exemple ici un dictionnaire { mot $\rightarrow$ étiquette }




## Intérêts

- Simplifier la tâche du modèle (améliorer les performances)
- Accélérer les traitements

# C'est parti !

- Taux d'erreur (en %) pour une tâche d'analyse morpho-syntaxique

| appr. | test | de | es | fr | id |
|---|---|---|---|---|---|
| ✗ | ✗ | 13.3 | 14.7 | 14.1 | 14.8 |
| ✗ | ✓ | 11.8 | 12.4 | 13.7 | 14.6 |
| ✓ | ✓ | 15.8 | 15.5 | 23.1 | 27.9 |

# C'est parti !

- Taux d'erreur (en %) pour une tâche d'analyse morpho-syntaxique

| appr. | test | de | es | fr | id |
|---|---|---|---|---|---|
| ✗ | ✗ | 13.3 | 14.7 | 14.1 | 14.8 |
| ✗ | ✓ | 11.8 | 12.4 | 13.7 | 14.6 |
| ✓ | ✓ | 15.8 | 15.5 | 23.1 | 27.9 |

## Double paradoxe

(a) Inclure des contraintes informatives pénalise le modèle

(b) Asymétrie lors de l'apprentissage et du test

# C'est parti !

- Taux d'erreur (en %) pour une tâche d'analyse morpho-syntaxique

| appr. | test | de | es | fr | id |
|---|---|---|---|---|---|
| ✗ | ✗ | 13.3 | 14.7 | 14.1 | 14.8 |
| ✗ | ✓ | 11.8 | 12.4 | 13.7 | 14.6 |
| ✓ | ✓ | 15.8 | 15.5 | 23.1 | 27.9 |

## Double paradoxe (apparent)

(a) Inclure des contraintes informatives pénalise le modèle

(b) Asymétrie lors de l'apprentissage et du test

# Première partie I

## Étude plus précise pour la tâche d'analyse morpho-syntaxique supervisée

# Formalisation

Soit une entrée donnée $\mathbf{x} \in \mathcal{X}$ et l'ensemble des sorties $\mathcal{Y}$.

## Fonction de contrainte

$c : \mathbf{x} \in \mathcal{X} \rightarrow \mathcal{Y}^c(\mathbf{x}) \subseteq \mathcal{Y}$

## Exemple

Soit $t : \mathcal{V} \rightarrow 2^{\mathcal{T}}$ un dictionnaire de type, on considère ici

$$t : \mathbf{x} = (x_1, \ldots, x_{|\mathbf{x}|}) \in \mathcal{X} \rightarrow \mathcal{Y}^t(\mathbf{x}) = t(x_1) \times t(x_2) \times \cdots \times t(x_{|\mathbf{x}|})$$

| Un | marché | pour | la | recherche | scientifique |
|---|---|---|---|---|---|
| DET | NOUN | ADP | DET | NOUN | NOUN |
| ADJ | VERB | NOUN | NOUN | VERB | ADJ |
| NOUN | | | PRON | | |
| PRON | | | | | |

# Modèle log-linéaire

Modélisation probabiliste

$$p_{\boldsymbol{\theta}}^{c}(\mathbf{y}|\mathbf{x}) = \frac{1}{\mathcal{Z}_{\boldsymbol{\theta}}^{c}(\mathbf{x})} \exp(\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}))$$

dans laquelle on normalise par

$$\mathcal{Z}_{\boldsymbol{\theta}}^{c}(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}^c(\mathbf{x})} \exp(\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}))$$

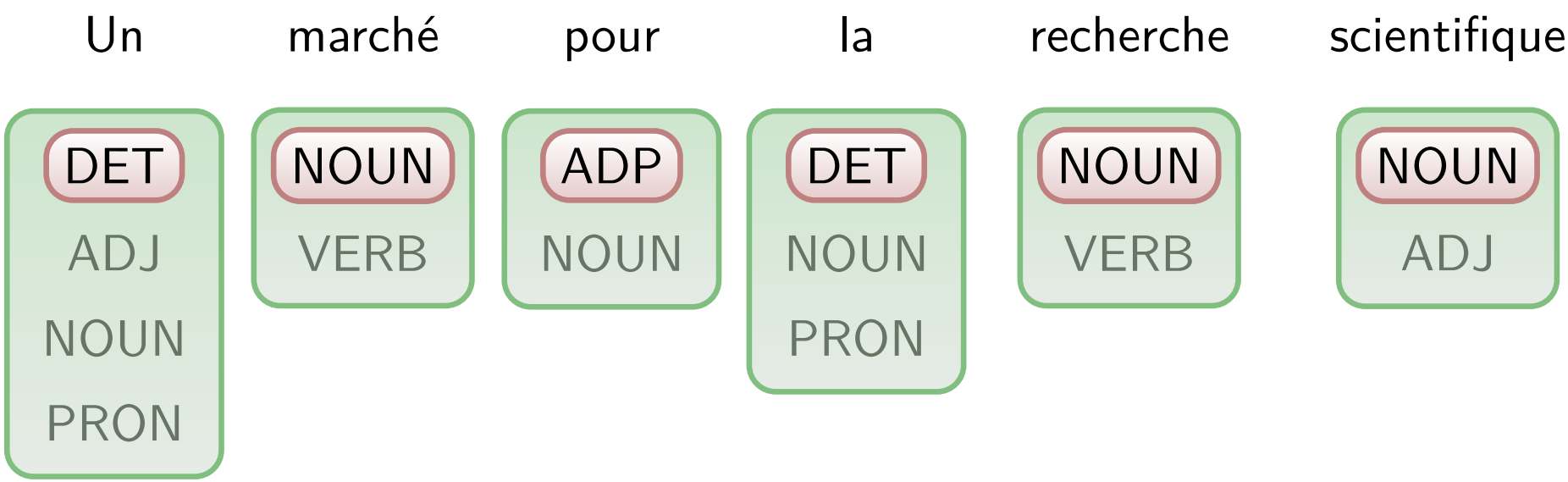
## Utilisation du maximum de vraisemblance

$$\arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log p_{\theta}^{c}(\mathbf{y}_i|\mathbf{x}_i) - \lambda_1 \|\boldsymbol{\theta}\|_1 - \frac{1}{2}\lambda_2 \|\boldsymbol{\theta}\|_2^2,$$

# CRF chaîne linéaire du premier ordre

Les caractéristiques se décomposent sur les paires d'étiquettes voisines

$$\boldsymbol{\phi}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{|\mathbf{x}|} \phi(y_i, y_{i-1}, \mathbf{x}).$$



# Cadre expérimental

- Analyse morpho-syntaxique de l'allemand
- Corpus arboré TIGER [Brants et al. 2004] (800k mots étiquetés)
- Tâche catégorie syntaxique seule : 54 étiquettes

  ADJA  ADJD  APPO  APPR  ...

- Tâche catégorie complète : 1373 étiquettes (619 observées)

  cs=ADJ, cas=gen, num=sg, gen=masc, pers=✗, tmp=✗, mode=✗

# Contraintes de type

## Corpus

Extraites sur le *corpus d'appentissage*

- Incorrectes et incomplètes

## Corrigées

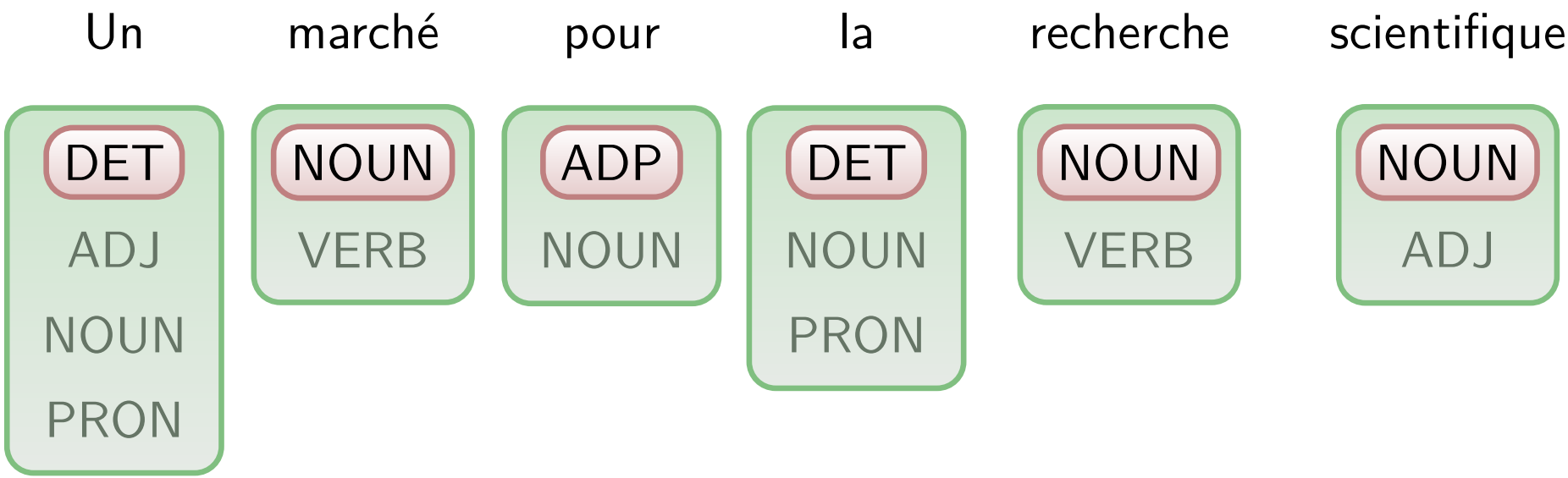
Corrigées pour les mots couverts en utilisant le *corpus de test*

## Oracle

Complétées pour tous les mots

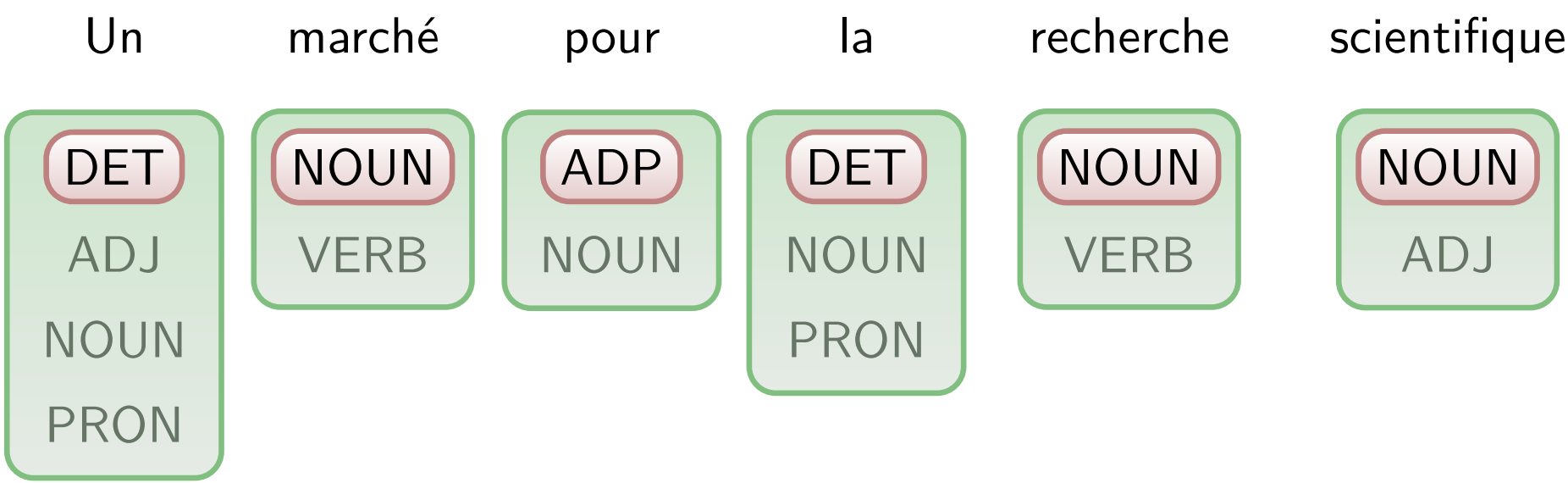
# Modèle MaxEnt simple

- Deux patrons de caractéristiques
  - (`mot`, `étiquette`)
  - (`étiquette`)



# Modèle MaxEnt simple

- Deux patrons de caractéristiques
  - (`mot`, `étiquette`)
  - (`étiquette`)



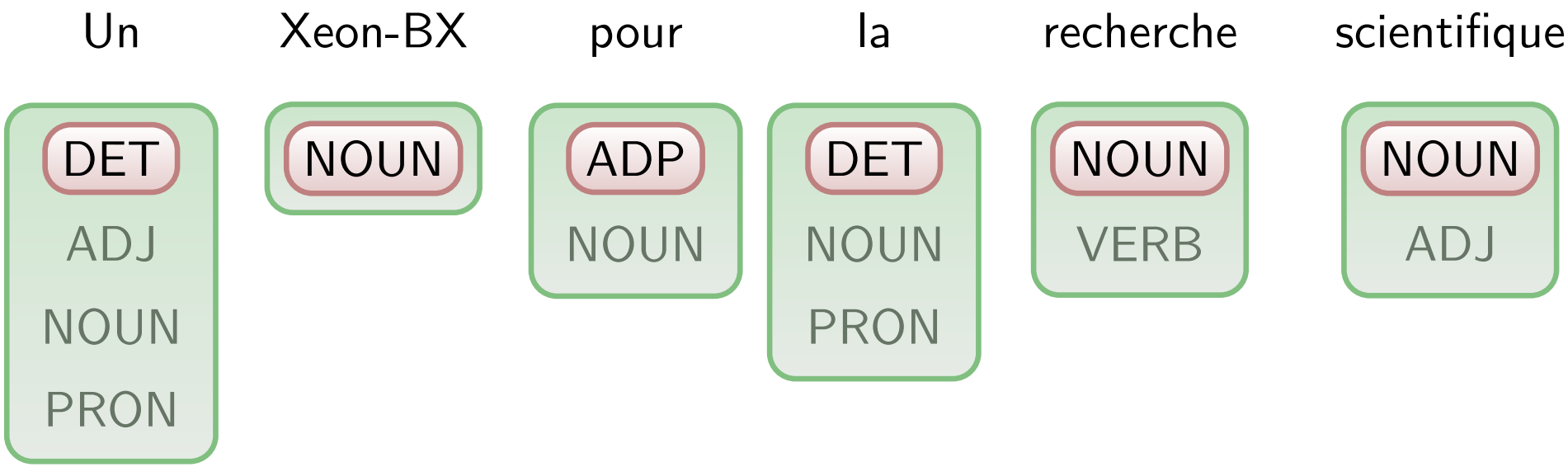

- Rem : les contraintes sont implicitement *encodées* dans le modèle

# Résultats (catégories syntaxiques seules)

| contraintes | | | MaxEnt | | | |
|---|---|---|---|---|---|---|
| type | appr. | test | global | MDV | MHV | amb |
| ✗ | ✗ | ✗ | 10.7 | 6.6 | 49.8 | 10.9 |
| corpus | ✗ | ✓ | 10.7 | 6.6 | 49.8 | 10.9 |
| | ✓ | ✓ | 15.5 | 6.6 | 100.0 | 11.0 |
| corr. | ✗ | ✓ | 10.7 | 6.6 | 49.8 | 10.9 |
| | ✓ | ✓ | 15.4 | 6.5 | 100.0 | 11.0 |
| oracle | ✗ | ✓ | 6.1 | 6.6 | 1.0 | 10.9 |
| | ✓ | ✓ | 6.0 | 6.5 | 1.1 | 11.0 |

# Commentaires

- Le problème vient des mots inconnus
- Problème possible : mots rares complètement désambiguïsés



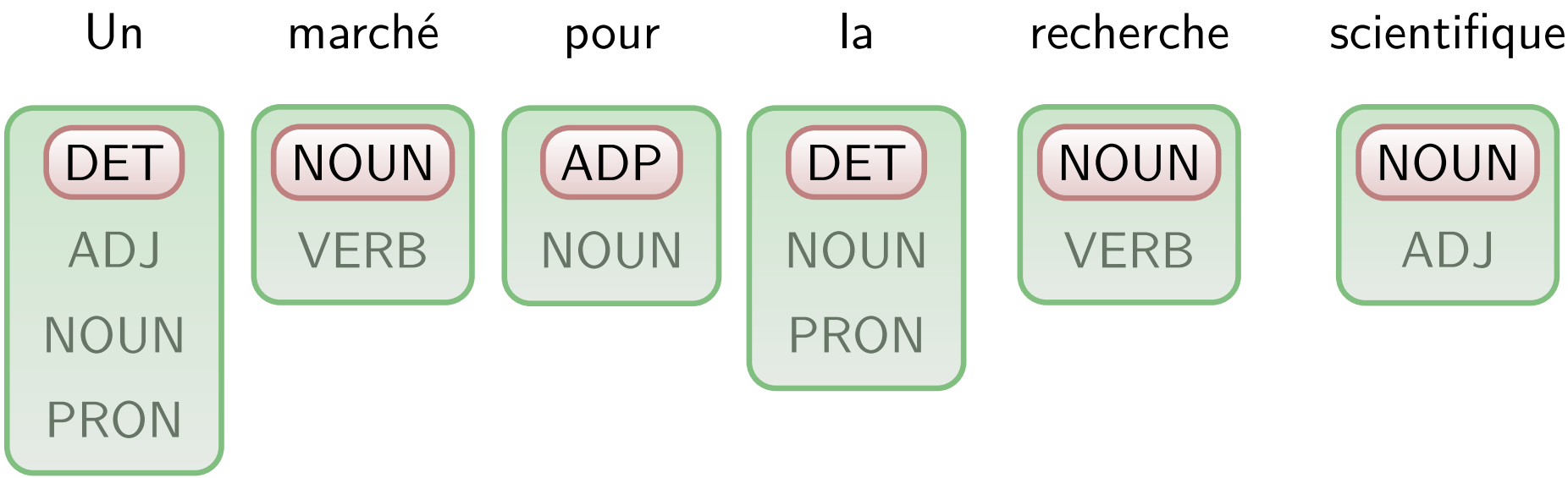

## Illustration

Étiquette la plus probable *sans* contraintes NOUN (50% des MHV)

Étiquette la plus probable *avec* contraintes APPRART (0% des MHV)

# Modèle CRF d'ordre 1

- Jeu de caractéristiques standard
  - Pour les mots courants, précédent et suivants
    - ★ mot en minuscules
    - ★ préfixes (jusqu'à taille 5)
    - ★ suffixes (jusqu'à taille 2)
    - ★ est en majuscule ou non
    - ★ contient un trait d'union
    - ★ est un nombre
    - ★ contient un chiffre
    - ★ sa forme avec et sans répétitions (e.g. Xxxx.Xxx et Xx.Xx)
    - ★ un biais



# Résultats (catégories syntaxiques seules)

| contraintes | | | CRF d'ordre 1 | | | |
|---|---|---|---|---|---|---|
| type | appr. | test | global | MDV | MHV | amb |
| ✗ | ✗ | ✗ | 2.9 | 2.2 | 9.4 | 3.0 |
| corpus | ✗ | ✓ | 2.9 | 2.3 | 8.8 | 3.0 |
| | ✓ | ✓ | 8.2 | 2.6 | 61.4 | 3.5 |
| corr. | ✗ | ✓ | 2.4 | 1.7 | 9.0 | 2.8 |
| | ✓ | ✓ | 7.7 | 2.1 | 61.6 | 3.4 |
| oracle | ✗ | ✓ | 1.6 | 1.7 | 0.4 | 2.8 |
| | ✓ | ✓ | 1.6 | 1.7 | 0.4 | 2.9 |

# Comment palier à ce problème ?

## Idée 1 : Mimer le comportement des mots rares à l'apprentissage

Mots de fréquence inférieur à un (hapax1), cinq (hapax5) ou dix (hapax10)

→ Ne pas utiliser les contraintes à *l'apprentissage* pour les mots rares.

- Problème : on est obligé de considérer *toutes* les étiquettes possibles

## Idée 2 : Assurer un minimum de contraste à chaque position

S'assurer qu'il y a à chaque position au moins $i$ compétiteurs.

→ Ajouter des étiquettes aléatoires là où ce n'est pas le cas.

- Variante : ne faire cela que pour les mots rares.

# Résultats (catégories syntaxiques seules)

| contraintes | global | MHV | amb |
|---|---|---|---|
| ✗ | 2.9 | 8.8 | 3.0 |
| hapax10 | 3.0 | 9.3 | 3.1 |
| hapax5 | 3.0 | 9.4 | 3.1 |
| hapax1 | 3.2 | 11.2 | 3.1 |
| min10 | 3.2 | 10.9 | 3.1 |
| min4 | 3.3 | 12.8 | 3.0 |
| min2 | 3.6 | 15.6 | 3.1 |
| corpus | 8.2 | 61.4 | 3.5 |

# Résultats (catégories morpho-syntaxiques complètes)

| contraintes | global | MHV | amb |
|---|---|---|---|
| ✗ | ? | ? | ? |
| hapax10 | ? | ? | ? |
| hapax5 | ? | ? | ? |
| hapax1 | 14.4 | 37.2 | 14.0 |
| min10 | 16.6 | 45.7 | 14.9 |
| min4 | 17.5 | 53.4 | 15.2 |
| min2 | 18.1 | 58.6 | 15.3 |
| corpus | 19.9 | 74.7 | 15.8 |

# Deuxième partie II

## Cas de l'analyse morpho-syntaxique faiblement supervisée

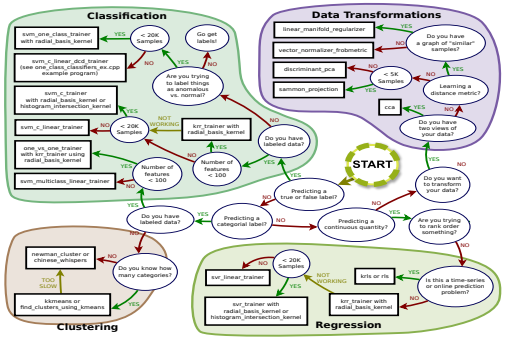
# Contexte

- Les méthodes supervisées nécessitent des annotations de référence manuelles
- Certaines langues sont moins bien dotées de telles ressources

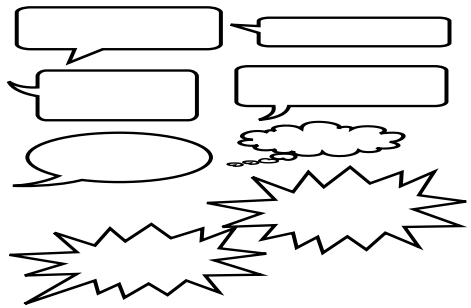
# Contexte

- Les méthodes supervisées nécessitent des annotations de référence manuelles
- Certaines langues sont moins bien dotées de telles ressources

- Une solution : le transfert cross-lingue



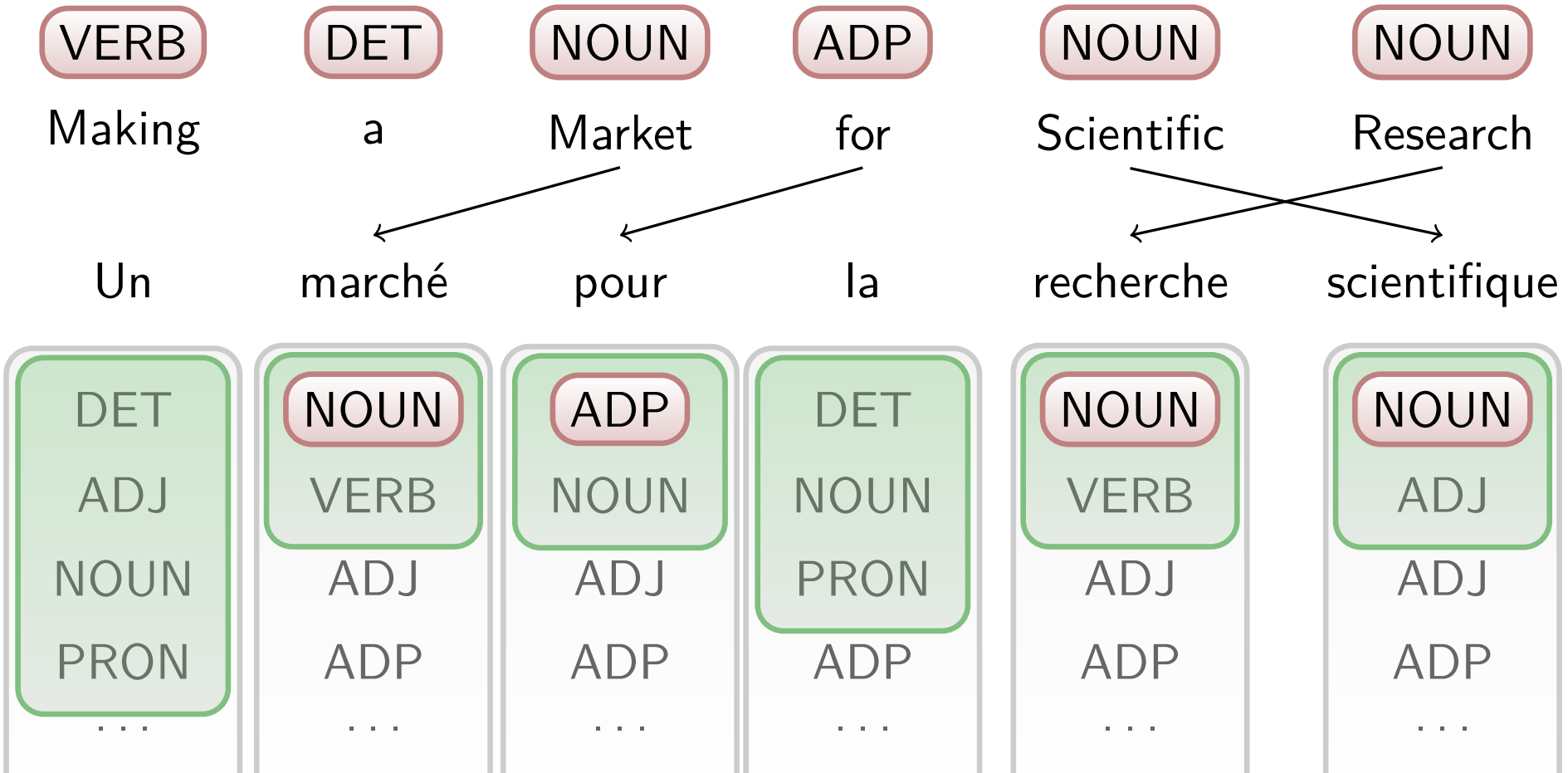

Transfer



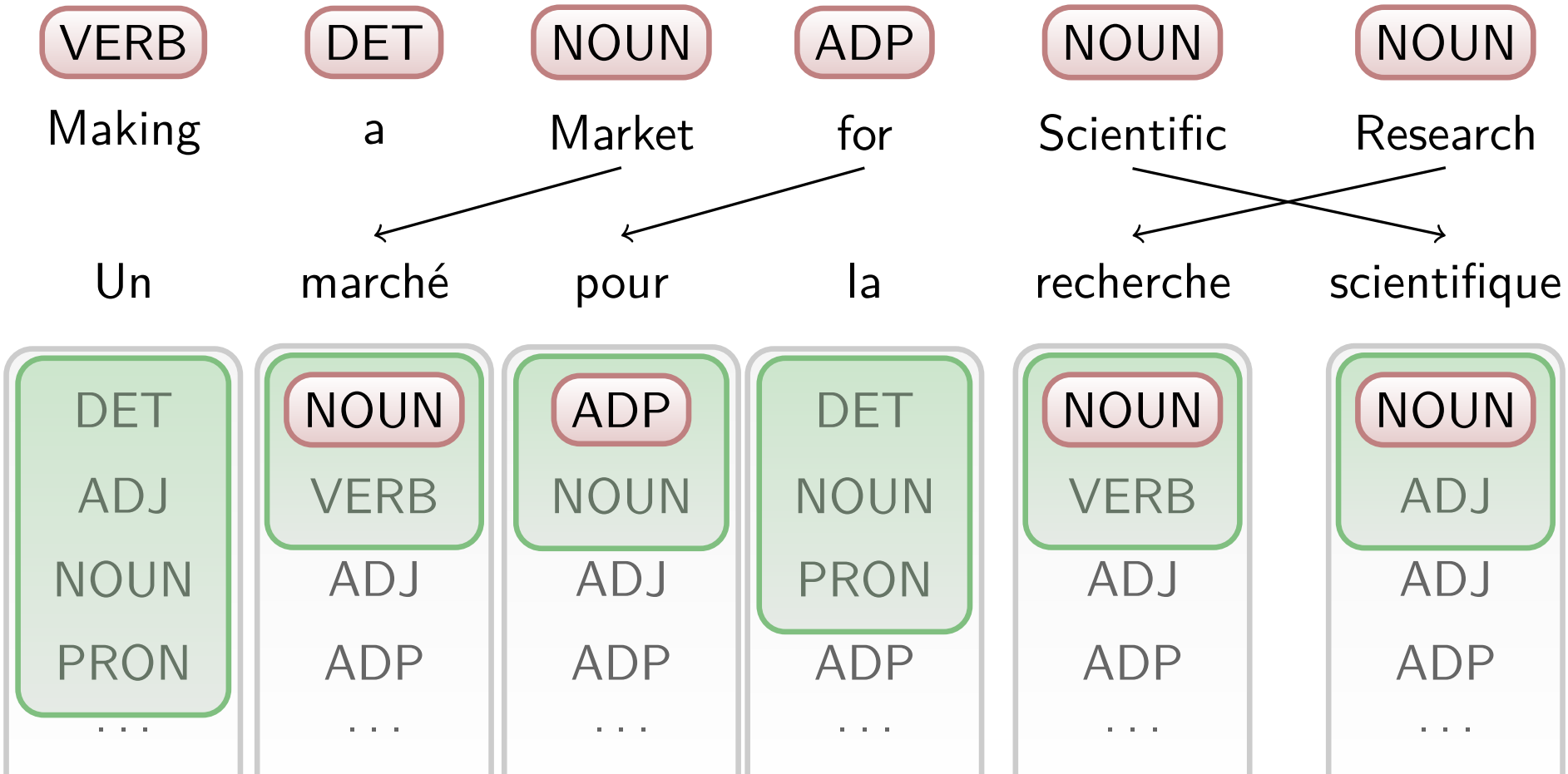
Ressource-rich language

Less-ressourced language

# Transfert cross-lingue d'annotations



# Transfert cross-lingue d'annotations




- Problème : annotations seulement *partielles*

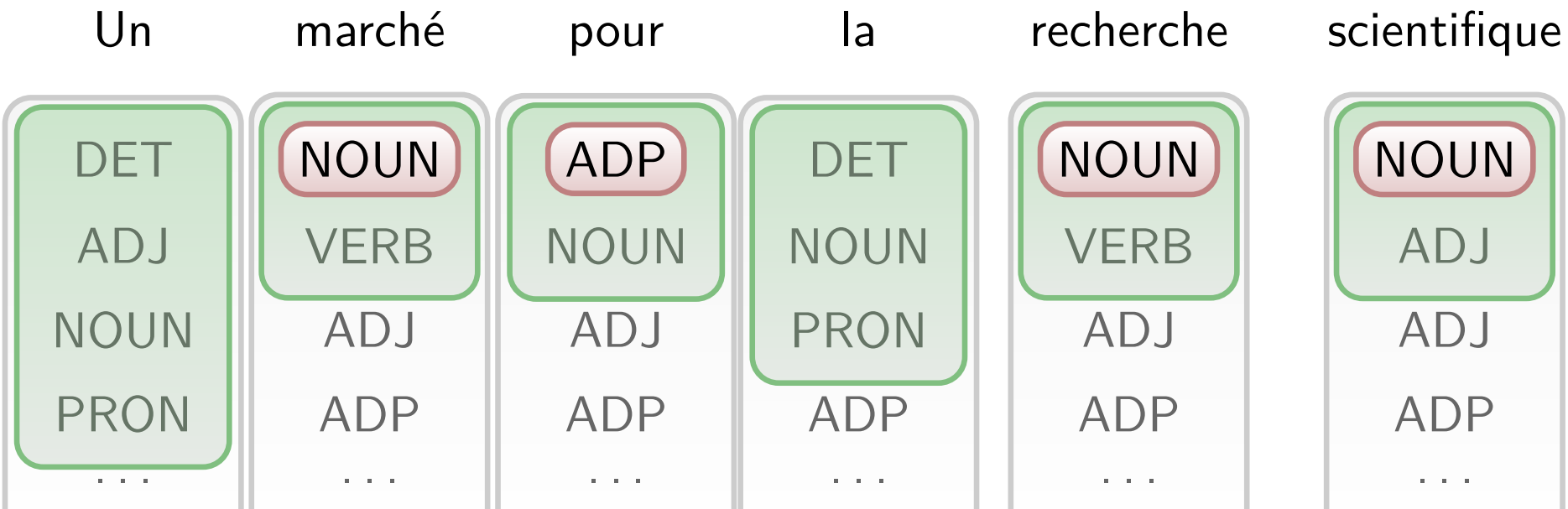
# Comment choisir les espaces de référence et de recherche ?

## Référence

Si il y transfert alors cette étiquette, sinon celles du dictionnaire

## Espace de recherche (à l'apprentissage et/ou au décodage)

- Soit les 12 étiquettes possibles
- Soit celles du dictionnaire
- Soit celles du dictionnaire si ceci est différent de la référence ($\otimes$)

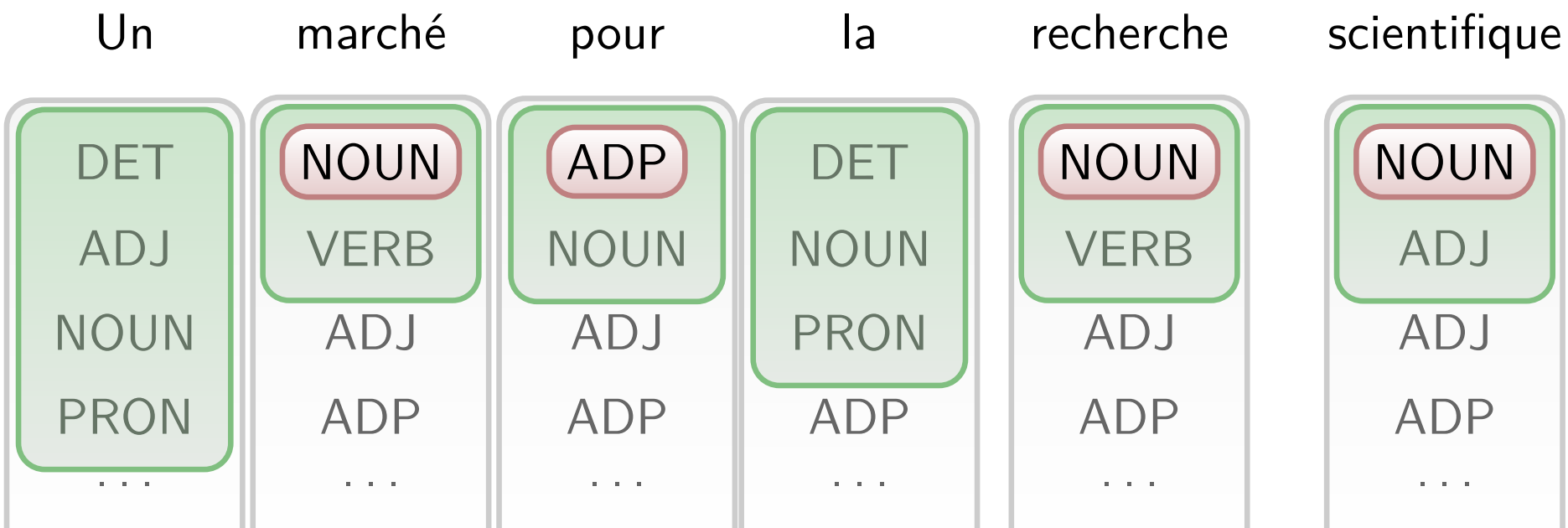


# CRF partiellement observé

$$\arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log p_{\theta}^{c}(\mathcal{Y}^{r}(\mathbf{x}_i)|\mathbf{x}_i) - \lambda_1 \|\boldsymbol{\theta}\|_1 - \frac{1}{2}\lambda_2 \|\boldsymbol{\theta}\|_2^2,$$

avec

$$p_{\theta}(\mathcal{Y}^{r}(\mathbf{x})|\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}^{r}(\mathbf{x})} p_{\theta}(\mathbf{y}|\mathbf{x})$$



# Cadre

## Dictionnaires

- Extrait automatiquement de Wiktionnaire
- Extrait du corpus d'apprentissage (après transfert)
- Combinaisons possible (union, intersection)

- On considère les 12 étiquettes universelles de Petrov et al. (2102).
- Expériences pour 9 langues de familles différentes
- Transfert depuis l'anglais (langue "riche")
- Corpus parallel : Europarl, NIST, Open Subtitle
- Pour plus de détails : [Wisniewski et al. 2014]

# Résultats

- Intersection de WIKTIONNAIRE et du dictionnaire issu du corpus

| appr. | test | cs | de | el | es | fi | fr | id | it | sv |
|---|---|---|---|---|---|---|---|---|---|---|
| ✗ | ✗ | 8.3 | 9.7 | 8.4 | 11.2 | 12.7 | 10.0 | 11.1 | 9.4 | 9.5 |
| ✗ | ✓ | 8.0 | 8.4 | 9.9 | 9.2 | 10.5 | 10.3 | 11.3 | 9.8 | 9.6 |
| ⊕ | ✓ | 8.0 | 8.8 | 12.6 | 9.3 | 11.4 | 11.9 | 11.9 | 10.8 | 9.7 |
| ✓ | ✓ | 12.8 | 13.2 | 14.0 | 12.0 | 22.4 | 14.7 | 20.5 | 14.7 | 14.6 |

# Résultats

- Intersection de WIKTIONNAIRE et du dictionnaire issu du corpus

| appr. | test | cs | de | el | es | fi | fr | id | it | sv |
|---|---|---|---|---|---|---|---|---|---|---|
| ✗ | ✗ | 8.3 | 9.7 | 8.4 | 11.2 | 12.7 | 10.0 | 11.1 | 9.4 | 9.5 |
| ✗ | ✓ | 8.0 | 8.4 | 9.9 | 9.2 | 10.5 | 10.3 | 11.3 | 9.8 | 9.6 |
| $\oplus$ | ✓ | 8.0 | 8.8 | 12.6 | 9.3 | 11.4 | 11.9 | 11.9 | 10.8 | 9.7 |
| ✓ | ✓ | 12.8 | 13.2 | 14.0 | 12.0 | 22.4 | 14.7 | 20.5 | 14.7 | 14.6 |

- Pourquoi $\otimes$ ne marche pas mieux ?

# Résultats

- Wiktionnaire seul

| appr. | test | cs | de | el | es | fi | fr | id | it | sv |
|---|---|---|---|---|---|---|---|---|---|---|
| ✗ | ✗ | 7.8 | 9.5 | 8.3 | 11.4 | 12.6 | 9.8 | 11.2 | 9.5 | 9.7 |
| ✗ | ✓ | 7.3 | 8.2 | 9.8 | 9.4 | 10.9 | 9.7 | 11.2 | 9.8 | 9.3 |
| ⊕ | ✓ | 7.3 | 9.0 | 14.5 | 9.8 | 11.4 | 9.6 | 12.2 | 12.4 | 9.6 |
| ✓ | ✓ | 8.8 | 10.7 | 16.9 | 10.3 | 12.1 | 10.9 | 13.9 | 13.4 | 10.1 |

# Résultats

- Wiktionnaire seul

| appr. | test | cs | de | el | es | fi | fr | id | it | sv |
|---|---|---|---|---|---|---|---|---|---|---|
| ✗ | ✗ | 7.8 | 9.5 | 8.3 | 11.4 | 12.6 | 9.8 | 11.2 | 9.5 | 9.7 |
| ✗ | ✓ | 7.3 | 8.2 | 9.8 | 9.4 | 10.9 | 9.7 | 11.2 | 9.8 | 9.3 |
| ⊕ | ✓ | 7.3 | 9.0 | 14.5 | 9.8 | 11.4 | 9.6 | 12.2 | 12.4 | 9.6 |
| ✓ | ✓ | 8.8 | 10.7 | 16.9 | 10.3 | 12.1 | 10.9 | 13.9 | 13.4 | 10.1 |

- Rem : on observe un comportement similaire avec le modèle à base d'historique de Wisniewski et al. (2014)

# Troisième partie III

Discutons !

# Comment intégrer des informations linguistiques ?

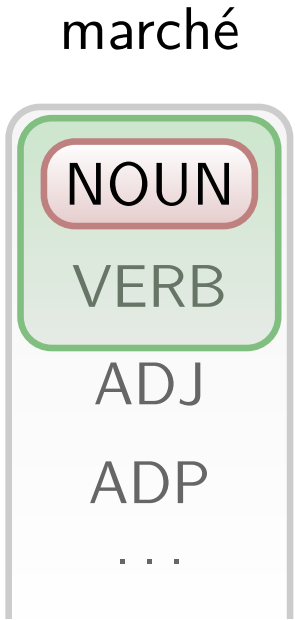
- Intégrer des connaissances linguistiques dans les modèles me semble un enjeu important du TAL
- Il existe d'autres moyens d'intégrer ces informations : sous forme de caractéristiques, par régularisation *a posteriori*, ou d'autres cadres d'apprentissage sous contraintes.

## Pistes de réflexion

- Il faut faire attention à la manière dont on intègre (même implicitement) nos connaissances
- Il faut encore que les connaissances ne soient pas "mieux apprenables" directement par le modèle.
- Comment mieux utiliser l'information dont on dispose ?

# Comment gérer des ensemble d'étiquettes trop grands ?

- À quoi sert vraiment "l'espace négatif" ?
- Peut-on en trouver une approximation moins coûteuse ?



# Un peu de philosophie

- À l'apprentissage, comment choisir les compétiteurs ?
- Comment régler le bon niveau de complexité pour un apprenant ?
- Quels sont les différentes réponses apportées en apprentissage statistiques et par les différents modèles ?

# Merci de votre attention


Questions ? Commentaires ? Suggestions ?