Entre corpus annoté et lexique sémantique : Quelles options pour le TALN ?

TALN'2000

Marc El-Bèze

Laboratoire Informatique d'Avignon

Désambiguïsation sémantique (WSD)
Pas une fin en soi !

- **Domaines d'application :**
  - Systèmes de dialogue, de Traduction Assistée, etc..
  - Pour la Recherche Documentaire (RD), idée de base :
    - Réduire le silence *expansion de requêtes par synonymie*
    - Réduire le bruit *traitement de la polysémie par WSD*
    - Pas d'orthogonalité entre modèles employés
    - **Þ résultats décevants**
  - Attentes fondées pour les moteurs de Q/A ?

La WSD a-t-elle un sens ?

- **Objectif réaliste ?**
  - **Peut-on atteindre un tel objectif ?**
    - Avec quel taux de performances ?
    - A quel niveau de granularité ?
  - **Légitimité de vouloir atteindre un tel objectif ?**
    - Lecture unique ? versus
    - Interprétations multiples
  - **Gare aux risques d'une WSD réductrice**

Lexique (pour humains)

- **But : Réduire une méconnaissance**
  - en s'appuyant sur du connu par le biais de
    - définitions, listes d'exemples
    - traits sémantiques, ...
- **Loupe déformante**
  - Accent mis sur ce qui est rare ***phrases complètes***
  - Mots usuels, moins bien traités ***formes prototypiques***
  - Mise en avant de sens anciens ***(désuets)***

Défauts des lexiques sémantiques

- **Lexiques trop ou pas assez spécifiques**
- **Entrées manquantes**
- adjectifs en -*able* en -*esque*
  - *gravable*, *caméléonesque*
- verbes en ré- ou en dé- :
  - déréguler, redénationaliser

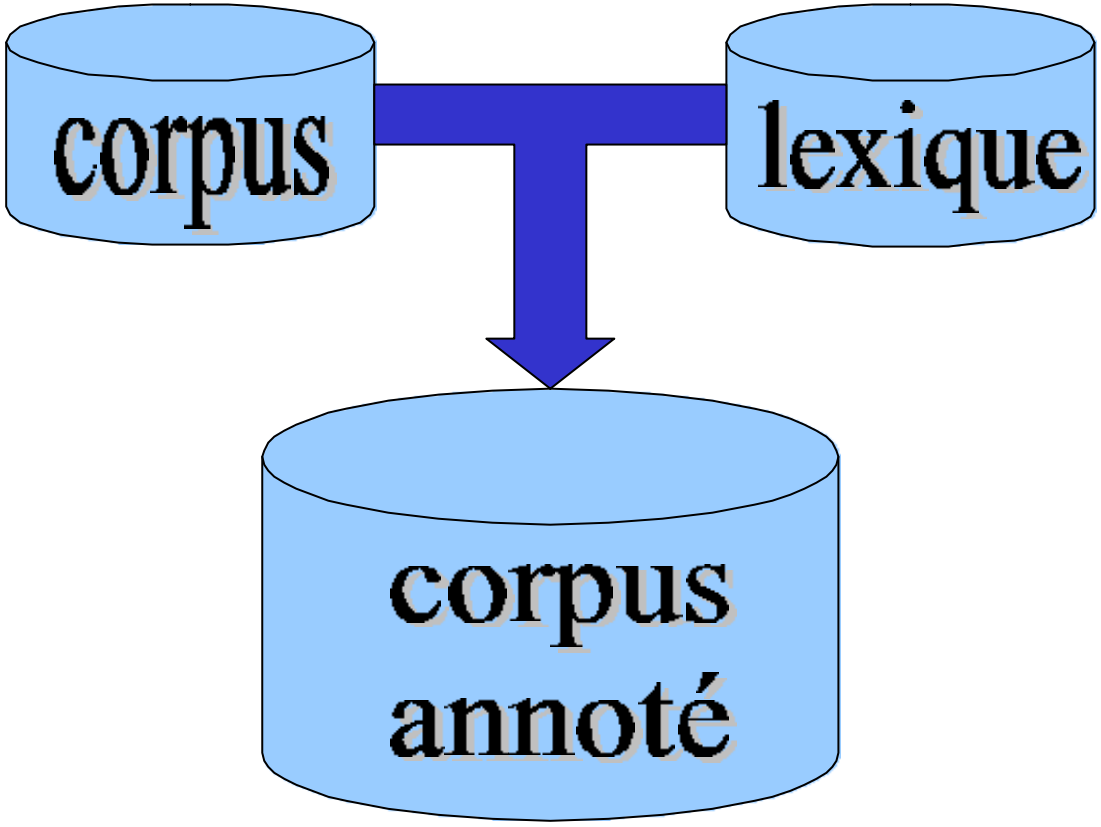
Découpage sémantique

- Manque de régularité des dictionnaires
  - alternance zones exhaustives / incomplètes
  - abondance de sens anciens
  - absence de sens plus récents
- Différences entre 2 dictionnaires
  - instructif de les observer sur des mots comme
    - raison, culture, lecture, ...

Listes d'exemples

- **Dictionnaire**
  - impossible d'être exhaustif
  - exemples bien choisis
    - ***« bien » pour êtres humains***
    - ***adéquation au TALN ?***
- **Corpus**
  - le plus grand possible ***jamais suffisant***
  - coût de l'annotation ***opposition avec le point précédent***

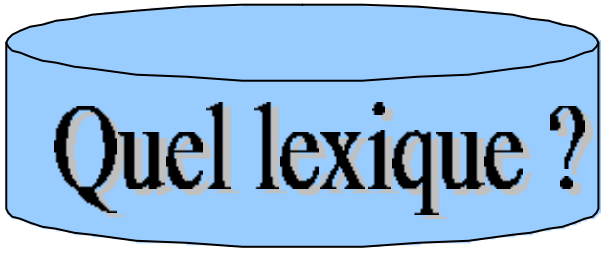
Quel lexique pour la WSD ? (I)



- **Liste d'étiquettes adaptée ?**

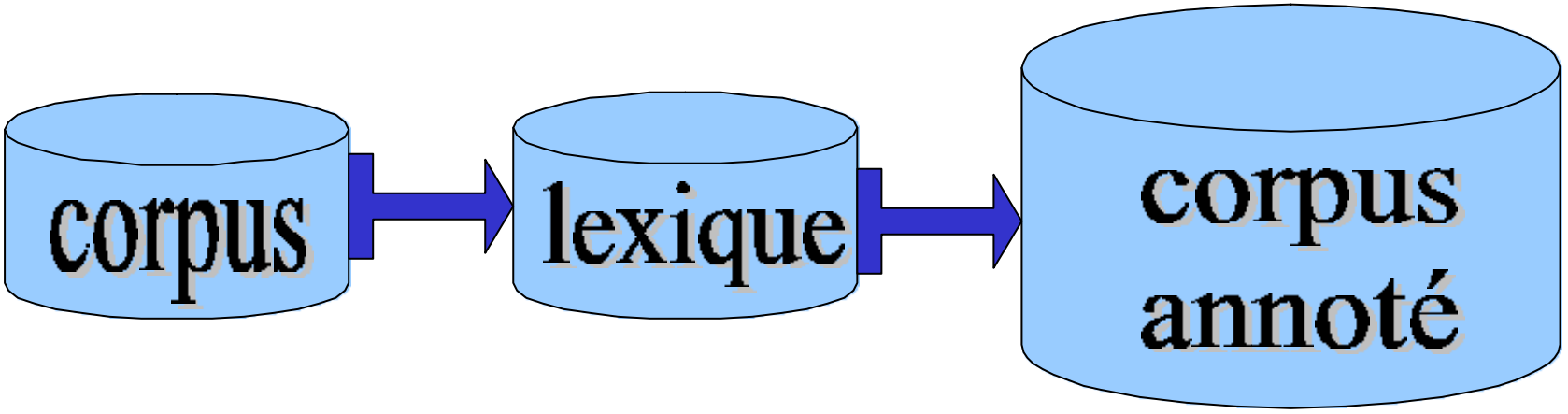
2 Problèmes et non 1

- **N. Ide & J. Véronis, CL 98 :**
- ***- Sense inventory is the problem n° 1 in WSD***



Quel lexique pour la WSD ? (II)

- **Une autre façon de procéder :**



Etude psycho-cognitique

- **Capacité d'êtres humains à relier**
  - définitions de mots
  - emplois de ces mots en contexte
- **Liste d'emplois : quelle liste de sens ?**
- **Résultats** ***(Julia Jorgenson 90? citée par C. Fellbaum)***
  - liste plus courte que celle d'un dictionnaire
  - ajout de contextes : la liste grandit très peu

Désambiguïsation

- Par contexte gauche même limité
  - *un livre / une livre / je livre*
- Mais que dire si pluriel ?
  - *les livres*
- Solution :
  - chercher le site informatif adéquat
  - s'appuyer sur le niveau linguistique approprié

Ambiguïtés syntaxiques / sémantiques

French word	Masculine noun	Feminine noun	Verb
coche	coach	nick, notch	nick, notch
greffe	clerk's office	graft	to graft
livre	book, register	pound	to deliver
manche	handle	sleeve, game, set	
moule	mould, matrix	mussel	to cast, mould
mousse	ship's boy	moss, lather	froam, lather
page	page-boy	page	
passe	master-key	passage, pasado	to pass, to go
perche	perch (fish)	perch, pole, rod	perch, roost
pique	spade (cards)	spite, quarrel	to bite, pique
poste	position, job	post-office	to mail
ponte	eminence	egg-laying	
tour	turn, trip, wheel	tower, castle (chess)	
voile	veil, cover, fog	sail, canvas	deam, shade

Sens et Sous-Sens : Exemples

- *La Poste*, une poste, je poste, le poste, ...
- c'est le désert / c'est un désert
- du monde, le monde, *Le Monde*
- **Autre exemple :** bureau
  - meuble, pièce, bureau d'une association ...
- Couvre-t-on tous les sens et sous-sens ?
  - aller au bureau, bureau de Poste, Bureau de Tabac, ...

Exemple senseval : *generous*

- **Définitions loin d'être exclusives**

  – **dico** 6 : étiquettes, **apprentissage** : 332 occurrences

*	104	**unstint**	sponsor, aide
*	26	**spacious**	large, ample
*	57	**kind**	bienveillant
*	87	**bigbucks**	somme + importante que prévue
*	46	**copious**	abondant
*	10	**liberal**	surestimation

Adéquation corpus / lexique

- 2 étiquettes *a* et *b*
- 4 cas se présentent :
  - une des 2 étiquettes, par exemple *a*
  - ni *a* ni *b*
  - *a* et *b*
  - ni *a* ni *b* mais entre *a* et *b*

Lexiques pour le TALN

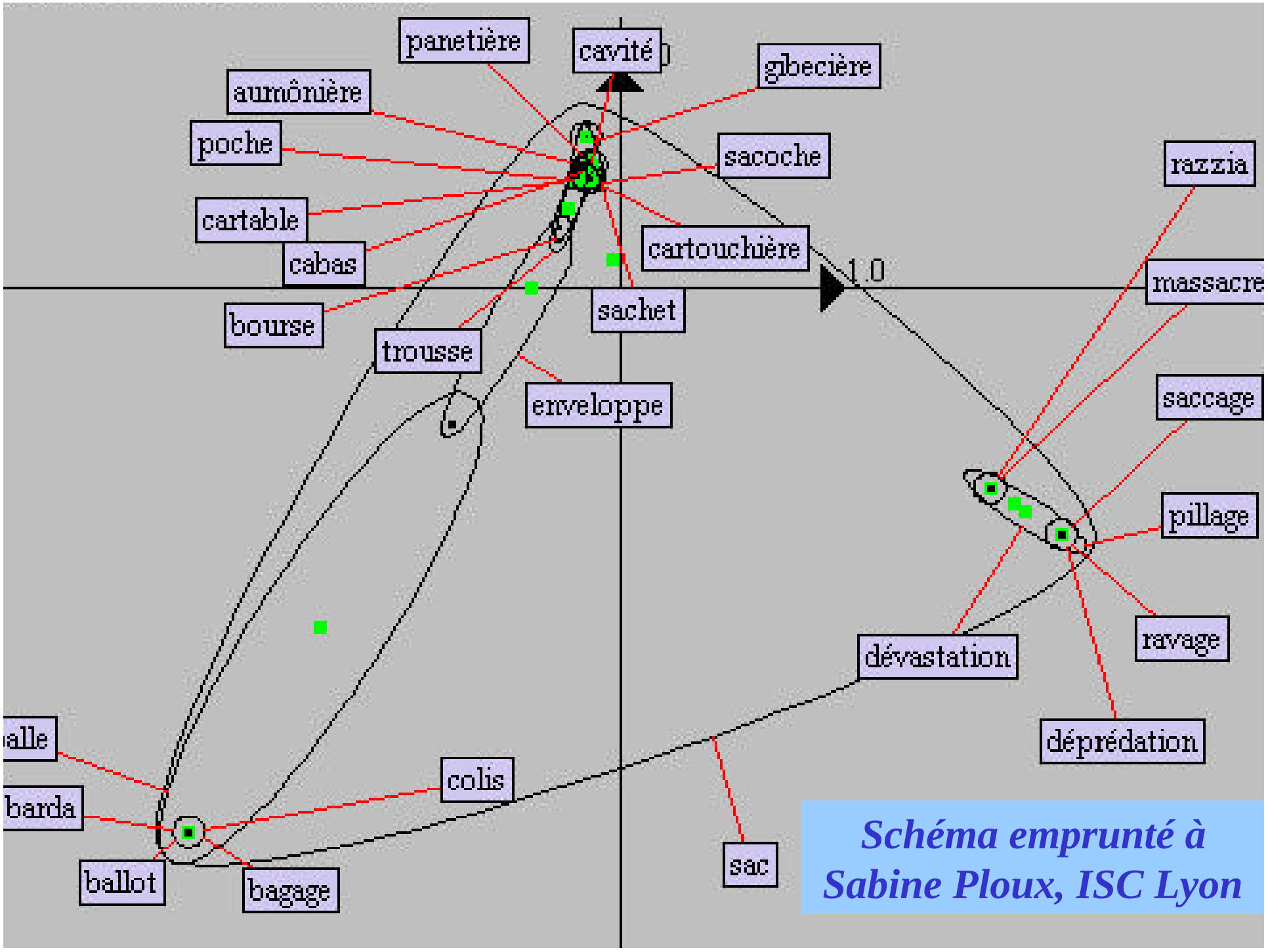
- **Attentes : optimiser la couverture**
  - sur le plan morphologique
  - sur le plan sémantique
- **Introduire du numérique dans un lexique**
  - fréquences d'usage
  - relations valuées
    - distances, indices de similarités,
    - degré d'appartenance à une classe

Fréquences & Lexique

- **Sous quelle forme ?**
  - qualitatif : inusité, rare, peu fréquent
  - ordinal : comme dans WordNet
  - décomptes ou probabilités
- **Contraintes** (pour les 2 derniers)
  - dépendance vis à vis d'un corpus

Relations valuées

- **Comment estimer l'intensité des relations ?**
  - A quel point 2 sens sont-ils
    - proches ou distants ? *quasi synonymie*
    - Cas particuliers l'un de l'autre ? *quasi hyponymie*
- **Comme pour les fréquences grâce à un corpus**
  - **utilisation des contextes**
    - globaux (thèmes, domaines, registres de langue)
    - courts $\Rightarrow$ *autres* classes (plus fines ? moins fines ?)


panetière
cavité
gibecière
aumônière
poche
sacoche
razzia
cartable
cartouchière
cabas
1.0
massacre
bourse
sachet
trousse
enveloppe
saccage
pillage
dévastation
ravage
déprédation
colis
barda
ballot
bagage
sac
Schéma emprunté à
Sabine Ploux, ISC Lyon

Evaluation des systèmes de WSD

- Quel protocole expérimental ?
  - Un ou plusieurs dictionnaires ?
  - Corpus multilingue ? ***Romanseval***
  - Evaluation sur un jeu limité de mots ? ***Senseval 1***
  - Evaluation sur tous les mots du corpus ? ***Senseval 2***
  - Niveau de granularité de l'évaluation ?

Problèmes ouverts

- Un ou plusieurs dictionnaires ?
  - Correspondance entre plusieurs dictionnaires
- Corpus multilingue ? (aligné ou non)
  - Lexique multilingue (cf. EuroWordnet)
- Evaluation sur un jeu limité de mots ?
  - comment classifier les contextes ?

Niveaux de granularité

- 3 niveaux dans Senseval : *coarse*, *middle*, *fine*
  - En fait 4 : corpus d'apprentissage : -*x*, -*p*, -*m*
- Disparités de niveaux possibles entre :
  - lexique / références du test / corpus d'apprentissage
- pourvu que ces différents niveaux correspondent
  - à une structuration hiérarchique des étiquettes

Comment gérer la hiérarchie ?

- ***Plus le grain est fin, plus la tâche est difficile***
- Prise de décision graduelle ?
  - niveau par niveau et
  - évaluation au grain le plus fin
- Ou au contraire ? Prise de décision globale
  - à un niveau plus fin que celui de l'évaluation

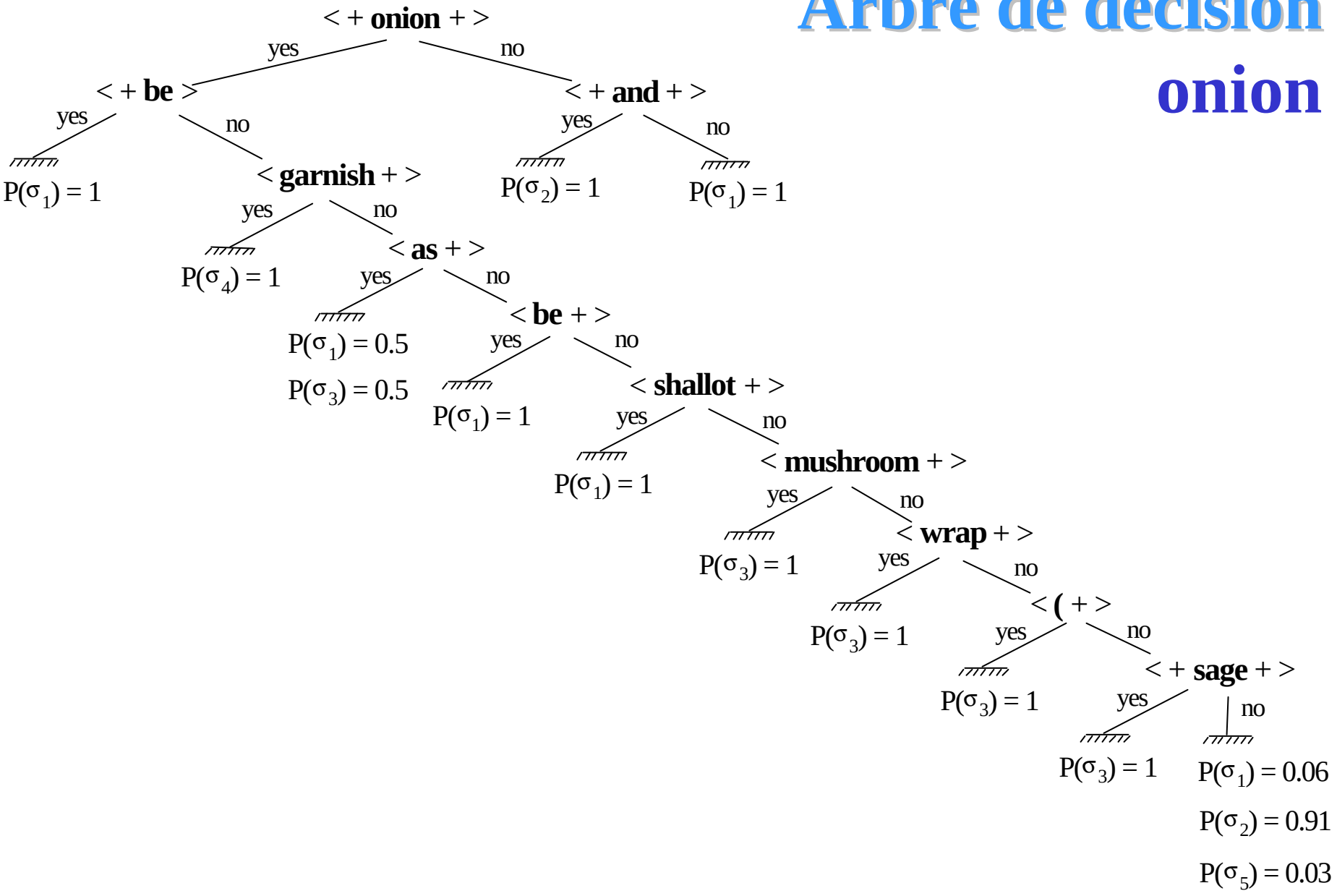
Hypothèses de travail

- Exploitation croisée **Dictionnaire & Corpus**
  - pour un mot donné *m*
    - un corpus d'exemples {*e*}
    - seul le mot *m* est étiqueté {*t*}
- But de l'apprentissage :
  - caractériser (***m, t***) au moyen de ***m {e} t***
    - similitude des (*m*,*e*,*t*), $\forall\, e$
    - différences entre (*m*,*t*) et (*m*,*t'*), $\forall\ t' \neq t$

Idée

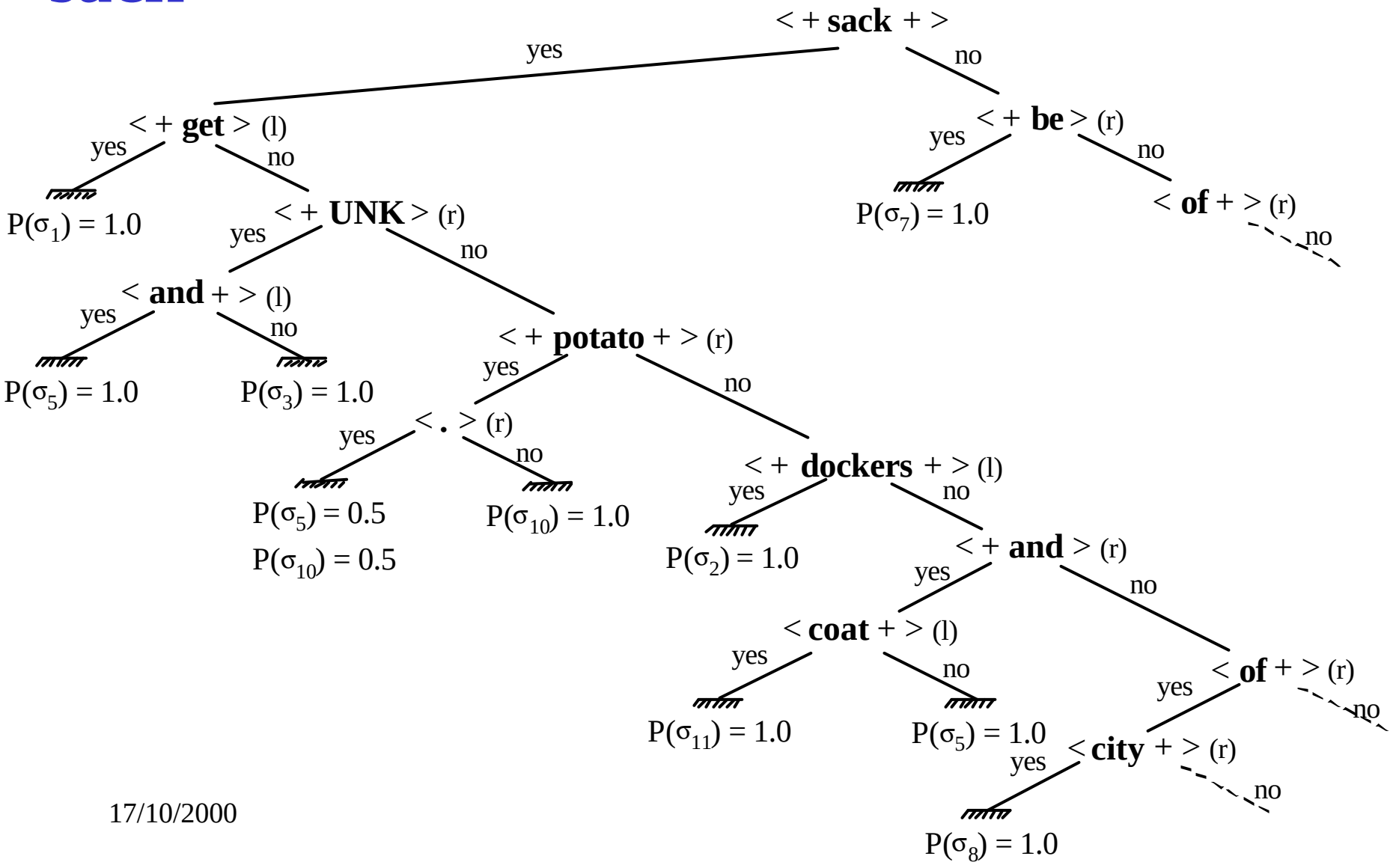
- **Pour généraliser différents contextes :**
- Appliquer une stratégie exploitant
  - les relations lexicales :
    - classes de synonymes, d'hyponymes, ...
    - traits sémantiques grossiers
    - classes syntaxiques, particularités graphiques
  - et une méthode de prise de décision
  - éventuellement des modèles

Arbre de décision onion

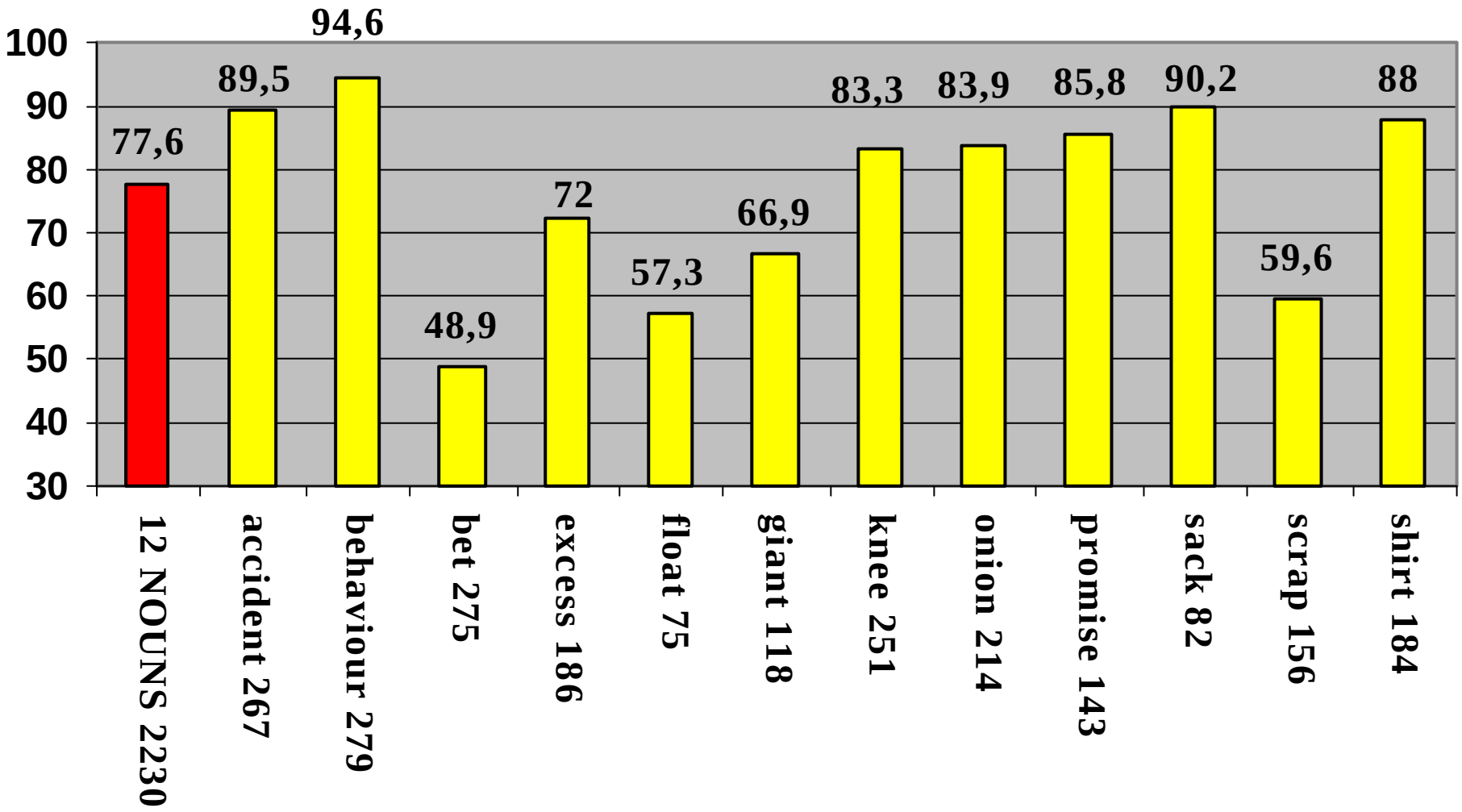


Arbre de décision

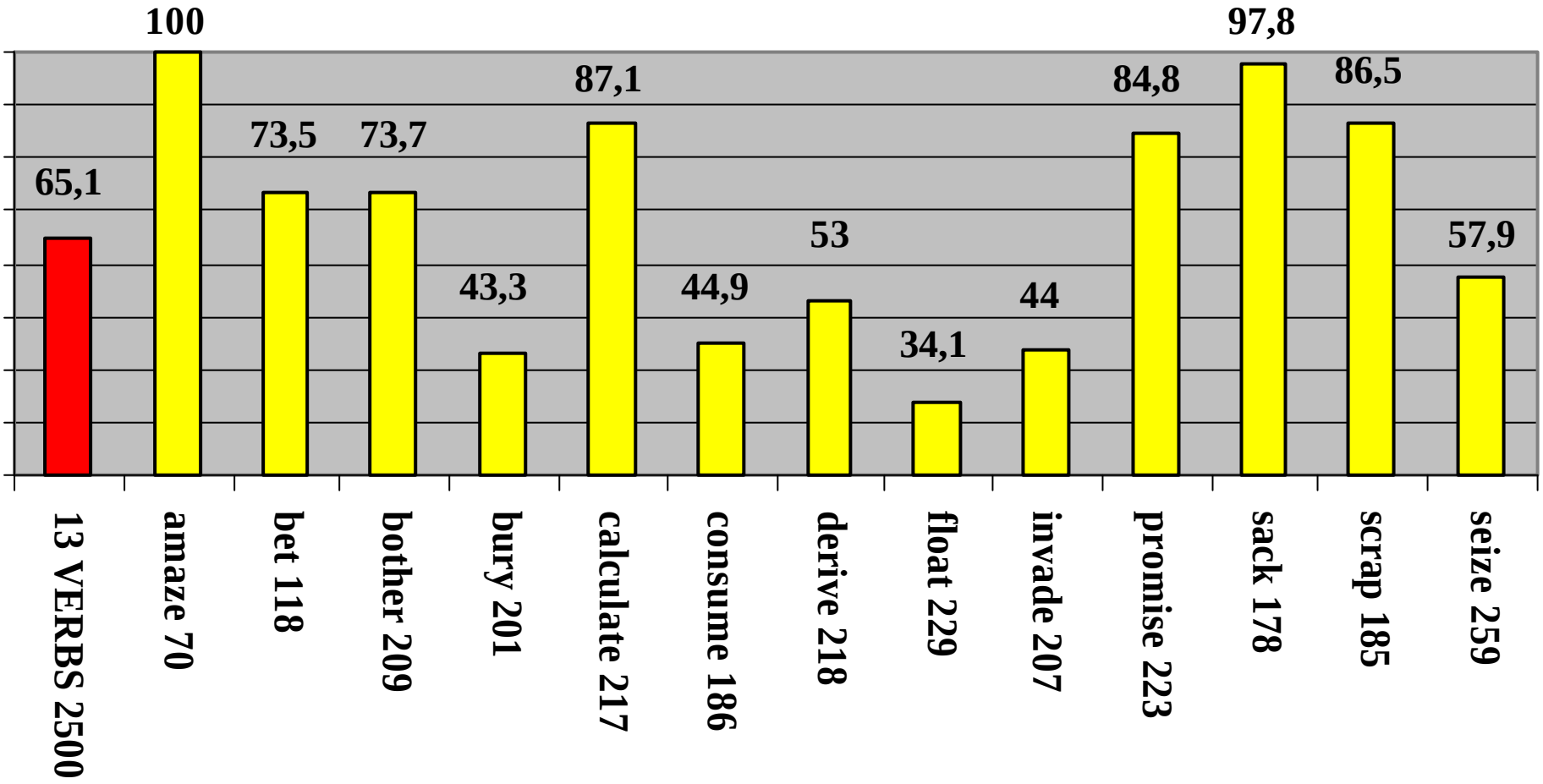
sack



12 Noms 2230 tests



13 Verbes 2500 tests



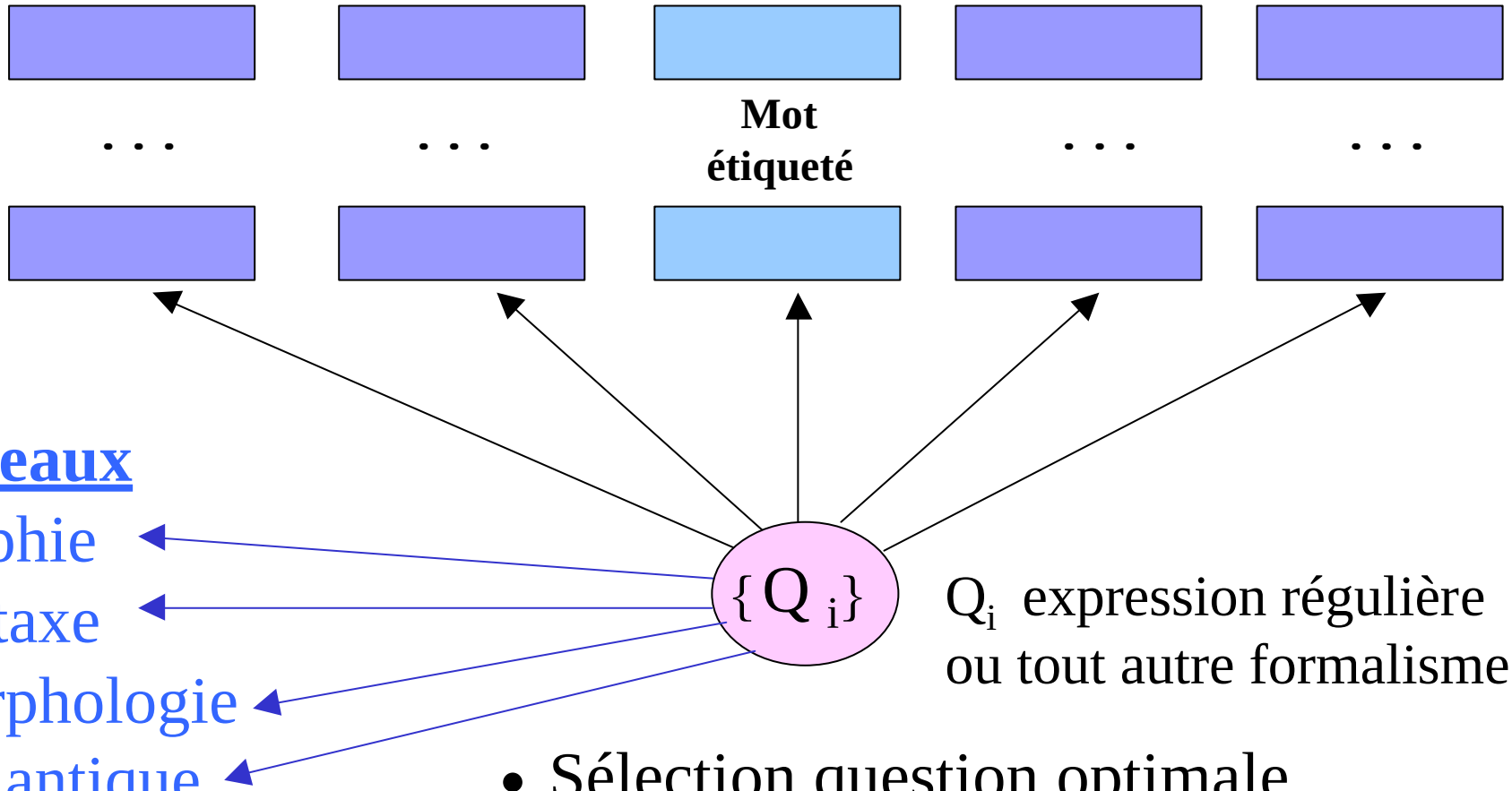
Corpus partiellement annoté (I)

- Evaluation sur un jeu limité de mots
- Corpus d'apprentissage insuffisant
  - problème bien connu : *sparse data problem*
- **Hypothèse :** entrées du lexique classifiées
  - selon un petit nombre de classes sémantiques

Corpus partiellement annoté (II)

- Utilisation des arbres de décision
- Questions portant sur ces classes
  - meilleure généralisation du contexte possible
- **Bénéfice secondaire**
  - réduction partielle de l'ambiguïté contextuelle

Approche préconisée MLSCT : Arbres de classification sémantiques multi-niveaux



- Sélection question optimale
- selon un critère donné
- (ex: réduction de l ’entropie)

Adaptation Etiquettes Corpus

- **Procédure itérative**
  - 1/ Jeu initial d'étiquettes + corpus annoté
  - 2/ Construction d'un arbre de décision
  - 3/ ACP sur les feuilles
  - 4/ Axes principaux nouvelles étiquettes
  - 5/ Aller en 2

CONCLUSION

- L'utilisation conjointe d'ACP et de SCT permet d'apporter une solution au problème de l'annotation sémantique d'un corpus avec un jeu d'étiquettes
  - obtenues automatiquement, adaptées au corpus
  - facilement interprétables
- **Þ Autre façon de voir un lexique sémantique**
- **Problème à résoudre :**
  - **Relations entretenues par les nouvelles étiquettes**

Références

- **CHUM 2000 Vol. 34 :**
  - **Special issue on SENSEVAL**
- **CL 1998 Vol. 24 N° 1 :**
  - **Special issue on WSD**
- ***WordNet & EuroWordNet***
  - ***Rapports techniques***
  - ***EuroWordNet A Multilingual DataBase with Lexical Semantic Networks, ed. Piek Vossen, Kluwer 98***