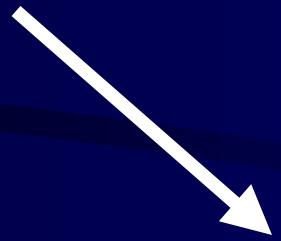
Évaluation de logiciels d'extraction de terminologie : examen de quelques critères

Marie-Claude L'Homme
Université de Montréal

- Plan
  - Unités recherchées et problèmes de base
  - Catégories de critères
  - Critères pré-évaluatifs (domaines d'application, stratégies d 'identification, langues traitées)
  - Performance : mesure du bruit (précision), mesure du silence (rappel)

Texte d’informatique



barre d’adresses
barre d’outils
barre d’outils personnalisables
bouton standard
chemin d’accès
menu approprié
mode d’affichage
nom de programme
outil de navigation
outil de navigation facultatif
page Web
volet d ’exploration

Accent mis sur l'extraction de syntagmes nominaux

Approche imparfaite dès le départ :

Pas de correspondance parfaite entre *syntagme nominal* et *unité terminologique*;
- certains termes ne sont pas des syntagmes nominaux
  (ex. *ordinateur*, *clavier*)
- certains syntagmes nominaux ne sont pas des termes
  (ex. *ordinateur neuf*)

Approches fondées sur les constats suivants :

La plupart des unités terminologiques sont complexes;

La plupart des termes sont de nature nominale;

Les termes sont répétés dans un texte spécialisé (Justeson et Katz 1995).

Problèmes de base (1 de 3)

1. Qu'est-ce qu'un terme ?

   Consensus difficile à trouver entre deux terminologues pour certaines unités terminologiques

   traitement automatique de la langue

   traitement automatique de l'écrit

   traitement automatique de l'oral

   traitement automatique de la parole

Problèmes de base (2 de 3)

2. Terme et non-terme : suites semblables; l'une est terminologique, l'autre non

ordinateur portable *ordinateur neuf*
poste de travail *utilisation des menus*
Il rapporta un vase de Chine
Il cassa un vase de Chine (Otman 1991)

3. Termes du domaine et termes appartenant à un autre domaine

termes d'informatique dans un texte comptable
termes de médecine dans un texte juridique
etc.
Textes d'informatique : 24 domaines (Love 2000)

Problèmes de base (3 de 3)

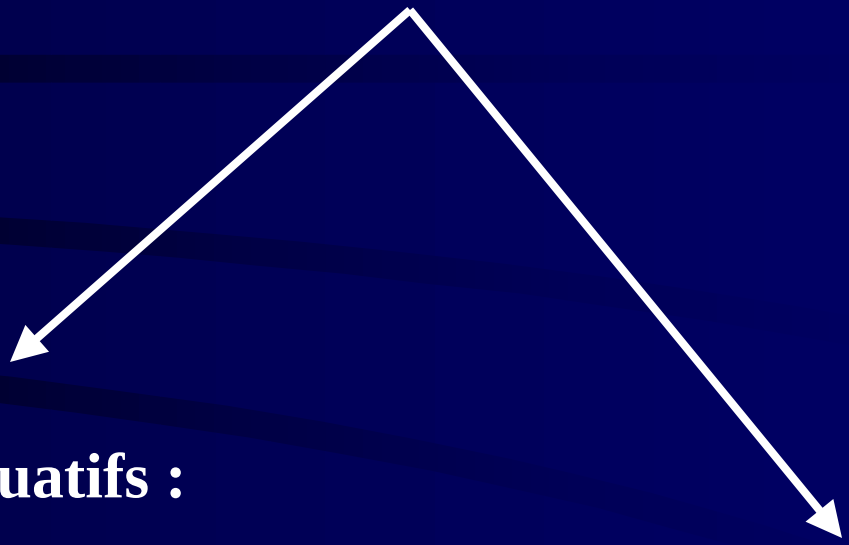
4. Découpage du syntagme nominal :

Les ensembles de systèmes de gestion de bases de données sont conçus pour permettre aux utilisateurs d'organiser, de gérer et de rechercher de l'information.

bases de données
ensemble de systèmes
ensemble de systèmes de gestion
ensemble de systèmes de gestion de base
ensemble de systèmes de gestion de bases de données
gestion de bases
gestion de bases de données
systèmes de gestion
systèmes de gestion de bases
systèmes de gestion de bases de données

ÉVALUATION

Deux catégories de critères



Pré-évaluatifs :

application

langue(s) traitée(s)

stratégies

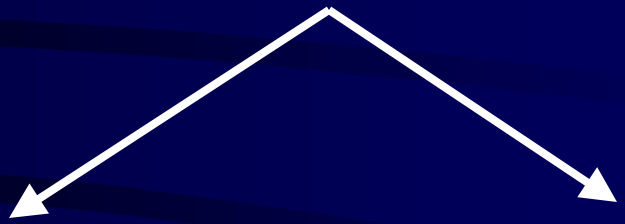
Performance :

bruit

silence

Critères pré-évaluatifs

Domaines d’application



Recherche d ’information (information retrieval)

Termes = descripteurs

Sujets abordés dans un document (classement par fréquence)

Applications « langagières »

Traduction : difficultés de traduction (la plupart sont de nature terminologique)

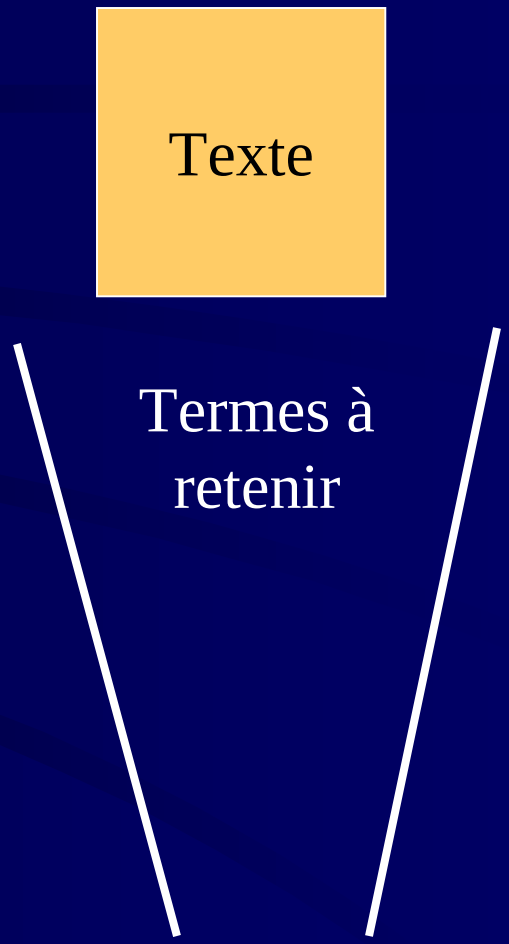
Terminologie : termes propres à un domaine à décrire

Critères pré-évaluatifs : Domaines d'application

Traduction : tous les termes se trouvant dans un texte à traduire (peu importe le domaine)

Terminologie : tous les termes se rattachant à un domaine de spécialité donné

« Nouveaux » termes : termes non documentés dans un répertoire terminologique : exclure ou signaler les termes qui s'y trouvent déjà. Nouveau dossier de traduction ou recherche de néologismes.

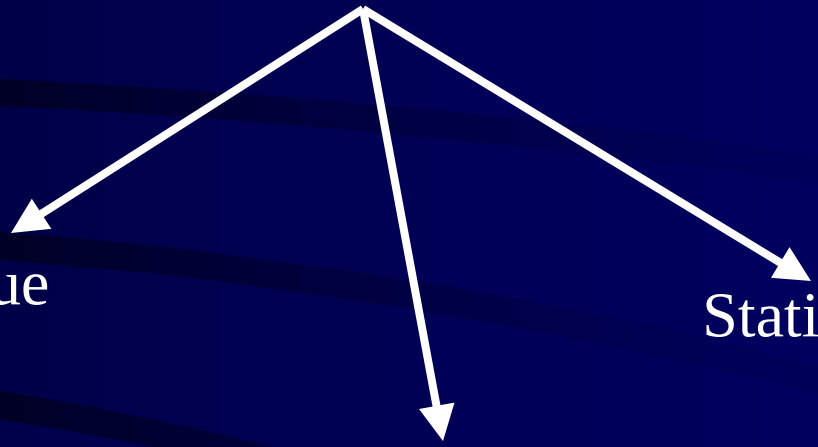


Critères pré-évaluatifs : Domaines d'application

Traduction → Intégration éventuelle dans un environnement de traduction assistée ou de traduction automatique :
base de données terminologiques;
logiciel de traduction automatique

Terminologie → Intégration éventuelle dans un environnement de terminologie assistée :
concordancier;
module de rédaction de fiches

Stratégies d'identification de termes



Linguistique

Statistique

Plus couramment : stratégies hybrides

statistique fondée sur des connaissances linguistiques;

linguistique avec prise en compte de la fréquence

1. Approche linguistique

Fondée sur le fait que les termes complexes se composent de suites de catégories grammaticales régulières :

Français
nom + adjectif : *menu déroulant*
nom + préposition + nom : *barre d'outils*
nom + préposition + verbe : *machine à laver*
nom + nom : *imprimante laser*
nom + préposition + déterminant + nom : *loi sur l'enfance*
toute combinaison plus longue des combinaisons ci-dessus :
nom + (nom + adjectif), nom + prép. + nom + prép. + nom, etc.

Critères pré-évaluatifs : stratégies d'identification de termes

1. Approche linguistique

la reconnaissance des catégories grammaticales repose sur :

1. le contenu d 'un dictionnaire

barre, nom
de, prép.
imprimante, nom
laser, nom
outil, nom

2. une analyse morphologique

analyse de la flexion et attribution d'une catégorie grammaticale

3. un texte étiqueté

Sélectionner/v./ *l'*/dét./ *imprimante* /n./ *dans* /prép./ *la* /dét./ *barre* /n./ *d'*/prép./ *outils* /n./.

1. Approche linguistique

a. Recherche de combinaisons de mots correspondant à des patrons

ex. nom + adjectif ; nom + prép. + nom (mettre dans la liste)
L'ordinateur portable est en général plus coûteux que l'ordinateur de bureau.

b. Coupes pratiquées dans le texte à partir d'éléments qui ne peuvent pas faire partie de termes (repérage de frontières)

ex. verbe, conjonction, préposition + possessif (Bourigault 1993)
Le circuit d 'aspersion de l 'enceinte de confinement / assure le / maintien / de sa / température nominale de fonctionnement / après une / augmentation de pression (Bourigault 1993).

2. Approches statistiques

Fondées sur le fait que les termes apparaissent fréquemment dans les textes spécialisés; des mots simples qui apparaissent fréquemment ensemble sont forcément significatifs

Exemple 1 : calcul de segments répétés

si des mots simples apparaissent plus de n fois ensemble dans un texte, la suite est extraite et placée dans la liste de termes complexes

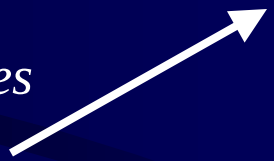
Exemple 2 : repérage de collocations

si un mot X apparaît plus fréquemment dans l’entourage d ’un mot Y qu’isolément, alors X et Y forment une collocation et cette dernière est placée dans la liste de termes complexes

2. Approches statistiques : filtrages nécessaires

Exemple en calcul des segments répétés

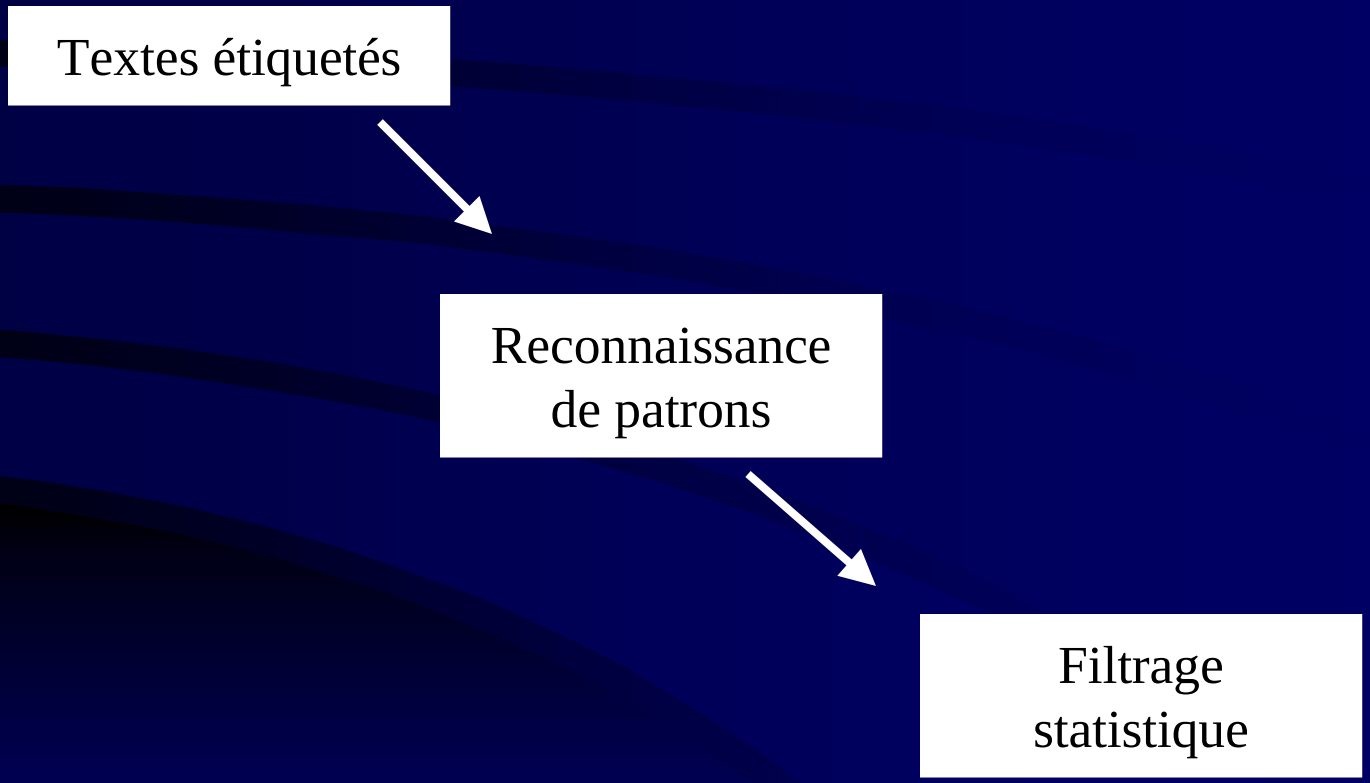
coupleur acoustique
la transmission
la transmission de
la transmission de données
le coupleur
le coupleur acoustique
transmission de
transmission de données



Exclusion de chaînes débutant ou se terminant par préposition ou un déterminant

coupleur acoustique
transmission de données

3. Stratégies hybrides : 1 exemple (Daille 1994)



Problèmes observés dans les approches linguistiques

a. dépendant de la langue traitée : appel à des connaissances linguistiques (dictionnaires, grammaires)

b. mots non répertoriés
correction : attribution d'une catégorie grammaticale par défaut

c. découpage du terme
stratégies : insertion de toutes les possibilités de découpage dans la liste ou uniquement la suite la plus longue

d. distinction termes et non-termes (ex. *matière volatile, mémoire volatile*)

e. patron inconnu (ex. *langage de très haut niveau*)

f. ambiguïtés (ex. *oil can; No oil can meet these requirements.*)
correction : analyse syntaxique (généralement locale)

Problèmes observés dans les approches statistiques

a. résultats plus intéressants sur des corpus de grande taille

b. non prise en compte de termes qui n’apparaissent qu’une seule fois dans le texte

c. découpage du terme
correction : calcul des syntagmes plus longs qui incluent un syntagme donné

d. distinction termes et non-termes
correction : calcul des fréquences des termes

plus grande attention accordée aux termes les plus fréquents dans certaines applications

Formes d’enrichissement de l’extraction terminologique

1. Lemmatisation

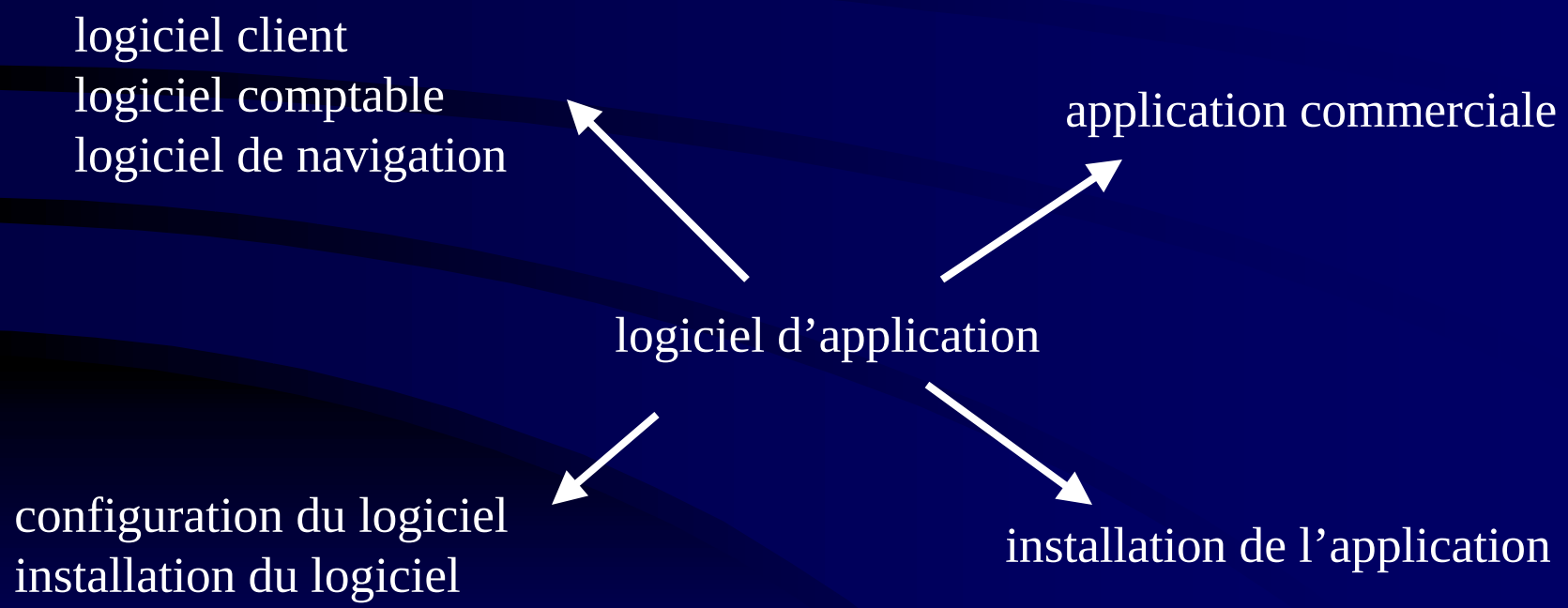
pression sanguine
pressions sanguines

pression sanguine	2
pression sanguin	2

*barre d’outils personnalisables

Formes d’enrichissement de l’extraction terminologique

2. Regroupement de termes par « familles »



Formes d'enrichissement de l'extraction terminologique

3. Regroupement des variantes terminologiques

a. deux termes

cellule du sang
cellule sanguine

b. variantes syntaxiques

professeur agrégé
professeurs adjoints et agrégés

Formes d’enrichissement de l’extraction terminologique

4. Identification de rapports sémantiques entre les termes

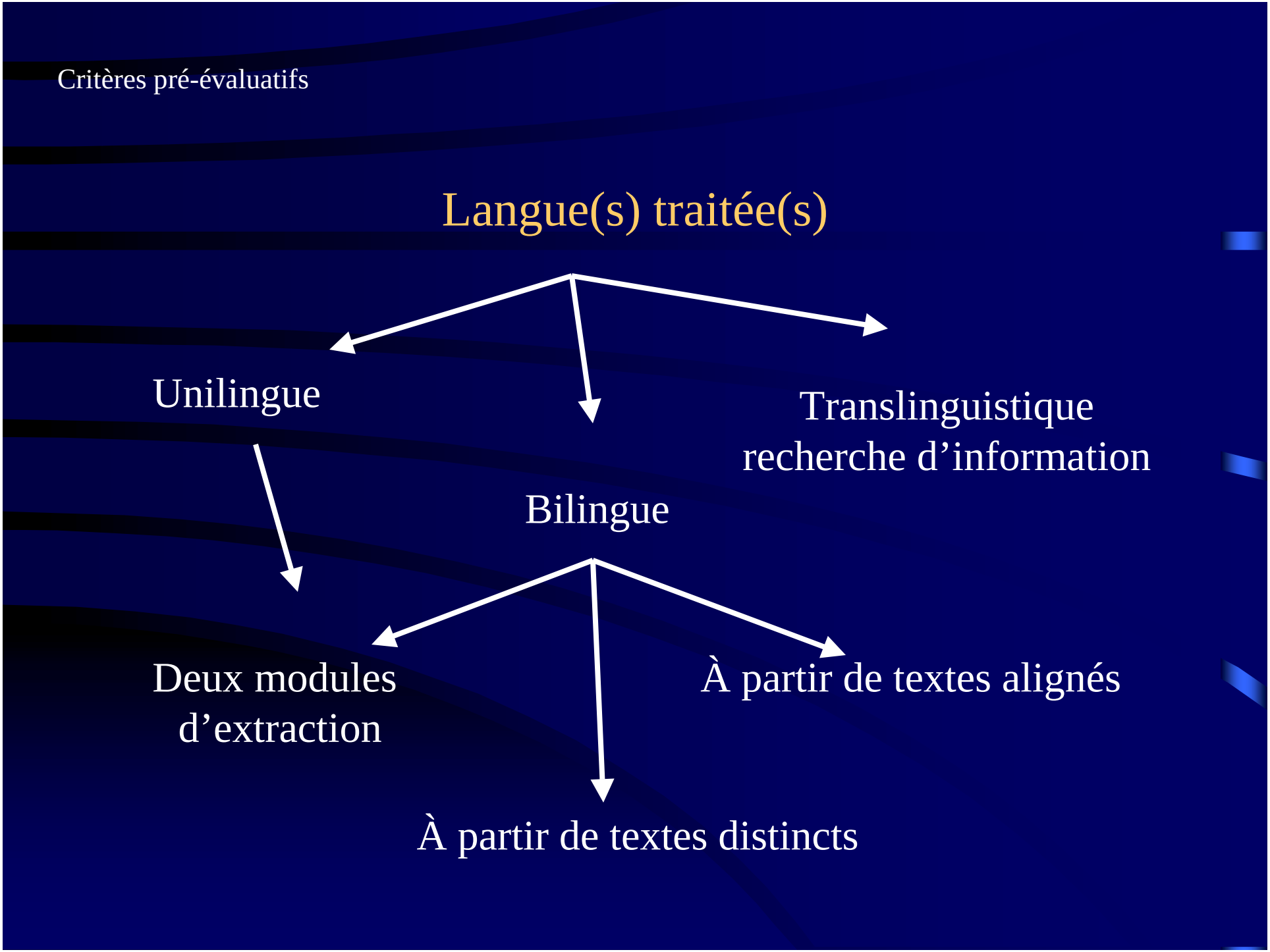
a. à partir de patrons linguistiques

x est un y : l ’imprimante laser est une imprimante

b. à partir des contextes dans lesquels apparaissent les termes : si deux termes apparaissent dans des contextes semblables, ils sont forcément liés sémantiquement

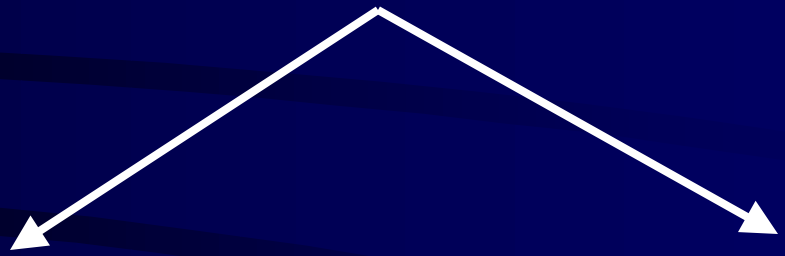
c. à l ’aide d’une ressource lexicographique externe (dictionnaire général ou spécialisé, thésaurus)

Critères pré-évaluatifs



Performance

Problèmes



Bruit : unité extraite mais pas pertinente du point de vue de l ’utilisateur

précision

Silence : unité pertinente présente dans le texte, mais pas extraite par le logiciel

rappel

Performance

Préférable de réduire le silence au maximum quitte à augmenter sensiblement la proportion de bruit;

Ordonnancement des termes par fréquence dans certaines applications (seuls les termes en tête de liste seront considérés)

Performance (concepteurs - utilisateurs)

Évaluation 1 : comparaison entre une extraction faite par l'humain et celle faite par le logiciel

comparaison des listes (silence, bruit)

Évaluation 2 : comparaison entre les extractions faites par deux logiciels différents

comparaison des listes (silence, bruit)

Évaluation 3 : comparaison du temps requis pour faire une extraction manuelle et une extraction automatique assortie d'une correction des problèmes

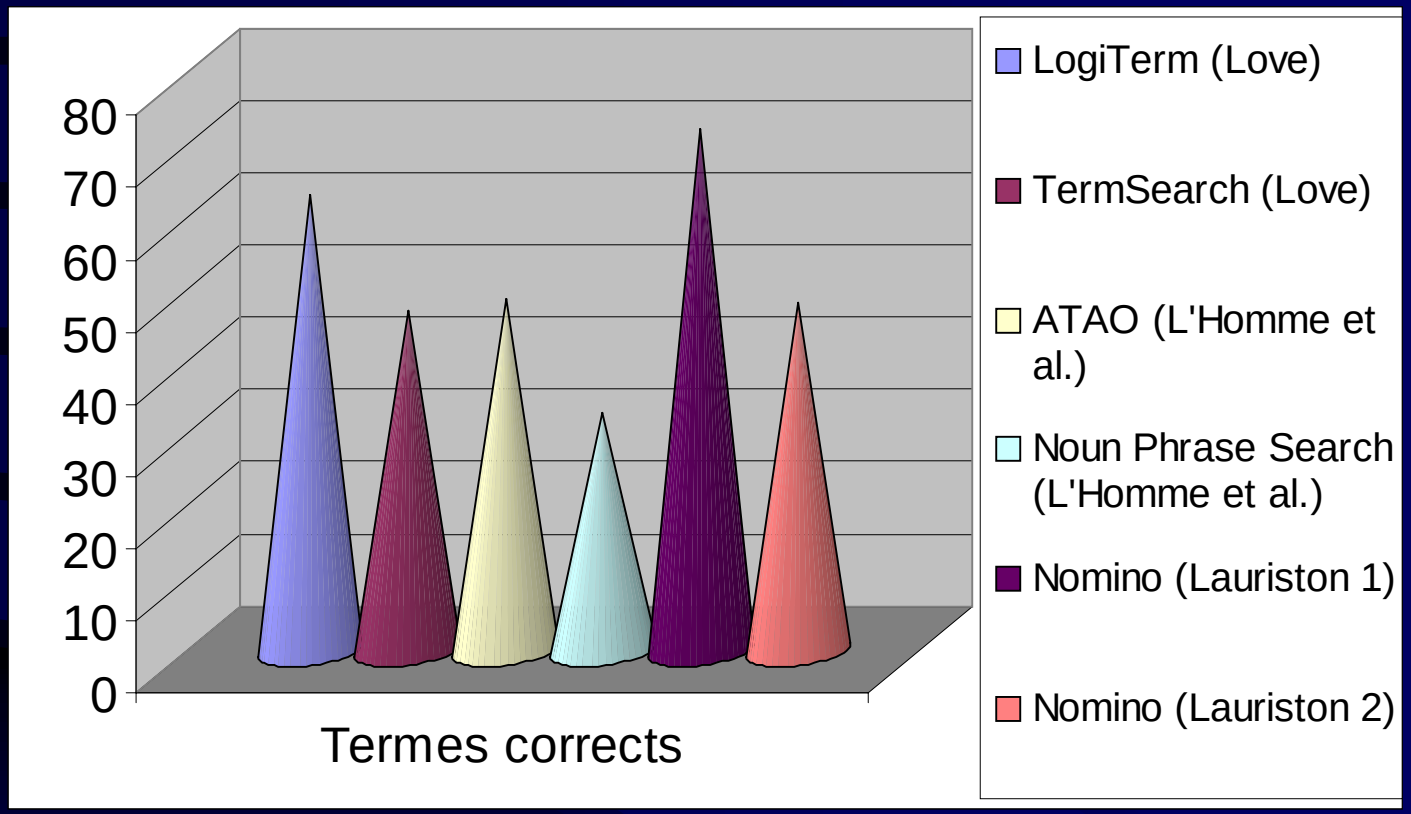
Performance

Méthodes :

1. Comparaison avec les entrées d ’un dictionnaire spécialisé : peu de termes présents dans les textes sont recensés dans les dictionnaires (Justeson et Katz 1995).

2. Comparaison avec une extraction humaine : identification des termes fondée sur l’intuition de l’extracteur humain (Love 2000).

Évaluation par différents auteurs



Conclusion

- Définir une application (recherche d’information, traduction ou terminologie);
- Distinguer les outils en fonction de la stratégie d’extraction afin de classer les erreurs;
- Tenir compte des enrichissements apportés aux outils;
- Distinguer les outils en fonction de la langue traitée;
- Évaluer en fonction d’une liste dressée par un humain;
- Soumettre les mêmes textes aux outils (tenir compte des types de textes);
- Classer les erreurs.