Compiler des dictionnaires spécialisés au moyen de techniques d'exploitation des corpus

Marie-Claude L'Homme
Observatoire de Linguistique Sens <-> Texte
Éclectik
Département de linguistique et de traduction
Université de Montréal

Point de vue adopté

Recherche terminographique dont l'objet est l'élaboration de dictionnaires spécialisés ou l'enrichissement de banques de terminologie à partir de corpus spécialisés.

textes spécialisés

Dictionnaire spécialisé

Banque de terminologie

Nouveaux types de dictionnaires spécialisés (1)

Address

Single code alloted to any workstation, computer or user connected to a network.

French base noun : adresse

Collocate noun + Base noun Element of an address (segment d'une adresse), extension of an address (suffixe d'une adresse), format of an address (format d'une adresse)

Collocate verb + Base noun To activate an address (activer une adresse), to assign an address (assigner une adresse), to attach an address (joindre une adresse)

Base noun + collocate verb Adress contains (adresse comporte), address links (adresse relie)

Collocate adjective + Base noun Absolute address (adresse absolue), dynamic address (adresse dynamique)

Meynard (2000)

Nouveaux types de dictionnaires spécialisés (2)

ARGENT DAFA (Binon et al. 2000)

- monnaie -- finance

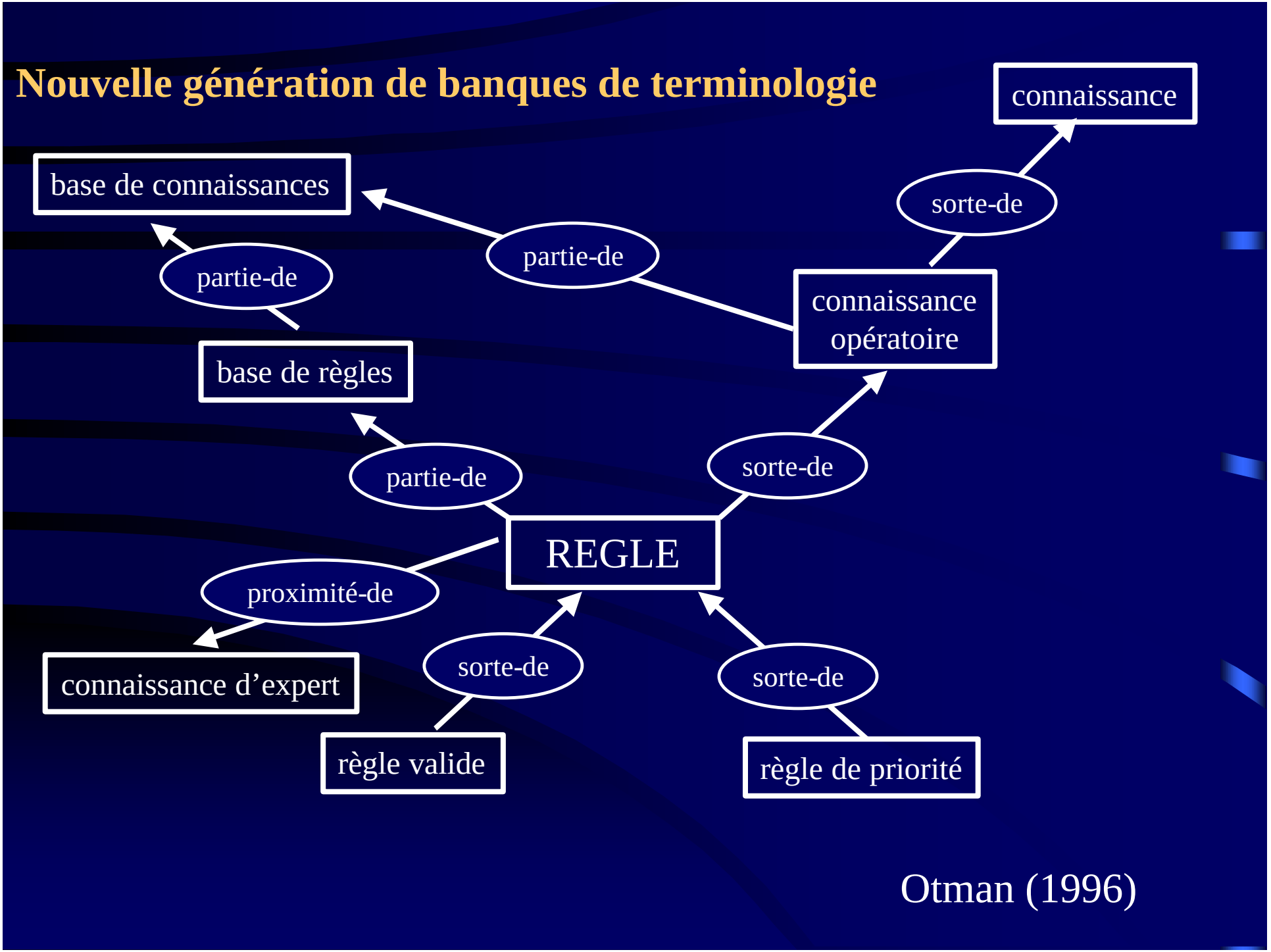
l'argent, un argentier, argenté, désargenté

1.1 Instrument de mesure de la valeur d'un bien…
1.2 Somme (importante) qui permet…
1.3 Métal précieux

être sans argent, être à court d'argent, l'argent ne fait pas le bonheur

argent liquide, argent comptant
somme (d'argent), valeur en argent
recevoir de l'argent, toucher de l'argent, gagner de l'argent

Nouvelle génération de banques de terminologie



Otman (1996)

Recherche terminologique

1. Collecte de la documentation
2. Repérage des termes
3. Collecte de données sur les termes
4. Analyse et synthèse des données recueillies
5. Consignation des données sur un support (ex. fiche)

Mise en forme d'un corpus en format électronique
Normalisation et conversion si nécessaire

Extraction de candidats-termes et autres unités terminologiquement pertinentes au moyen de logiciels d'extraction

Recherche et extraction de contextes contenant de l'information utile au moyen de concordanciers

Consignation des données sur support électronique (SGBD, logiciels de terminologie, SGML, XML)
Organisation des données terminologiques

Les données recherchées (1)

Attestation des termes : réalisation concrète et preuve de leur utilisation par des spécialistes et un ensemble de contextes

L'extrasystole ventriculaire se définit comme une dépolarisation ventriculaire prématurée, non précédée par (ou sans relation chronologique fixe avec) une onde P, le ventriculogramme étant large (supérieur à 0,12 s.), et déformé avec aspect soit de bloc de branche droit s'il s'agit d'une extrasystole ventriculaire naissant du ventricule gauche, soit de bloc de branche gauche s'il s'agit d'une extrasystole ventriculaire naissant du ventricule droit. La repolarisation est inversée avec une onde T négative. Le diagnostic d'extrasystole ventriculaire est le plus souvent simple sur l'électrocardiogramme de surface, le seul "piège" étant une extrasystole supra-ventriculaire avec bloc de branche fonctionnel.

Les données recherchées (2)

Données sur la fréquence et la répartition des termes

Dans un corpus médical de 500 000 mots

	occurrences	répartition
extrasystole ventriculaire :	92	12
dépolarisation ventriculaire	8	4
ventriculogramme	13	9
électrocardiogramme	113	14

Les données recherchées (3)

Des variantes terminologiques

Le présent chapitre est consacré à l'étude d'un système d'exploitation bien précis: le *DOS* (*Disc Operating System*).

Mais dire que les droits et libertés de la personne sont inséparables du bien-être général ((...))

Elles se trouveraient implicitement circonscrites par les autres libertés et droits de la personne.

((...)) préambule, affirmant que les droits et libertés de la personne humaine sont inséparables des droits et libertés d'autrui et du bien-être général.

la collectivité ne peut souffrir de l'exercice des droits et libertés individuels et que, par conséquent, ces derniers trouvent leurs limites non seulement là où commencent les droits et libertés d'autrui, mais également lorsque leur exercice compromettrait l'intérêt collectif.

Les données recherchées (4)

Des régularités formelles

**ose* dans un corpus médical

acidose
arthrose
cyphose
drépanocytose
hémosidérose
ostéophytose
ostéoporose
resténose
scoliose
sténose
thrombose
xanthomatose
etc.

imprimante dans un corpus d'informatique

imprimante à 9 aiguilles
imprimante à 24 aiguilles
imprimante à bande
imprimante à impact
imprimante à jet d'encre
imprimante à laser
imprimante à marguerite
imprimante à marteaux
imprimante matricielle
imprimante par caractère
imprimante par points
imprimante sans impact
imprimante thermique
etc.

Les données recherchées (5)

Des indices sur d'autres relations sémantiques

a) Séries morphologiques

Programme, programmer, programmable, programmation, programmeur, reprogrammer, etc.

Compiler, compilation, compilable, recompiler, etc.

Compatible, compatibilité

b) Liens sémantiques entre termes formellement apparentés ou non

Étiqueteuse, étiquetage
Étiquette, étiquetage
salle de vente, enchère
centre commercial, locomotive
boutique satellite, satellite
Article, produit
Article, assortiment

Les données recherchées (6)

Des éléments définitoires

L'extrasystole ventriculaire se définit comme une dépolarisation ventriculaire prématurée, non précédée par (ou sans relation chronologique fixe avec) une onde P, ((...))

Elle affecte la rétine -- le tissu nerveux au fond de l'oeil qui transmet les messages visuels au cerveau.

Le DOS permet de: · faire fonctionner le matériel qui compose votre système microinformatique: imprimante, unité de disques, clavier, modem ou tout autre périphérique;· configurer votre matériel:

Les données recherchées (7)

Des indices sur les liens conceptuels entre termes

Vis de transmission et son hyperonyme à savoir *pièce*

La vis de transmission est une pièce de machine utilisée pour transformer un mouvement de rotation en un mouvement de translation ((...))

Disque virtuel et son holonyme, c'est-à-dire *mémoire*

Un disque virtuel est une partie de la mémoire configurée comme une disquette. ZMSPC2

Moteur à piston et ses méronymes, à savoir *bloc de fonte, bloc d'alliage léger*

Un moteur à piston est constitué d'un bloc de fonte ou d'alliage léger percé de trous qui forment les cylindres. Dans ceux-ci, coulissent les pistons raccordés par les bielles au vilebrequin.

Les données recherchées (8)

Des cooccurrents : les unités lexicales se combinant de façon privilégiée avec les termes

térêt de l'exploration vers une	maladie	locale. Cette radiographie des sinu
le d'éviter l'évolution vers la	maladie	postphlébitique. La strycardiomyopa
le d'éviter l'évolution vers la	maladie	postphlébitique. La streptokinase e
L'adrénoleucodystrophie est une	maladie	récessive liée au sexe qui débute e
omalie et/ou trouble du rythme;	maladie	rythmique de l'oreillette. Dans la
ellement l'adulte alors que les	maladies	dysmyélinisantes se rencontrent hab
ellement l'adulte alors que les	maladies	dysmyélinisantes se rencontrent hab
Les leucodystrophies sont des	maladies	métaboliques d'origine enzymatique
. Les leucodystrophies sont des	maladies	métaboliques d'origine enzymatique
t mitral, insuffisance mitrale,	maladies	mitrales, prolapsus valvulaire mitr
criniennes, il faut évoquer les	maladies	osseuses constitutionnelles dont la

Le corpus : extraction de contextes à partir de chaînes de caractères

TERME RECHERCHÉ : maladie

L'adrénoleucodystrophie est une	maladie	récessive liée au sexe qui débute en
e enzymatique responsable de la	maladie	le diagnostic peut être fait plus ou
méningo-encéphalites liées à la	maladie	de Lyme. Les anomalies découvertes da
t à l'expression clinique :- la	maladie	de LETTERER-SIWE : atteinte granuloma
t plus fréquents au cours de la	maladie	de Recklinghausen que dans la STB, l'
rale est rare au cours de cette	maladie	, car les tumeurs les plus classiques
térêt de l'exploration vers une	maladie	locale. Cette radiographie des sinus

Le corpus : extraction de contextes à partir d'étiquettes morphosyntaxiques (1)

Les <le; dét.; masc. plur.> étiquettes <étiquette; nom; fém. plur.> peuvent <pouvoir; verbe; ind. prés. 3e pers. plur.> varier <varier; verbe; inf. prés.> en <en; prép.> fonction <fonction; nom; fém. sing.> de <de; prép.> l' <le; dét.; fém. sing.> application <application; nom; fém. sing.> visée <viser; verbe; part. passé; fém. sing.>, mais <mais; conjonction> elles <il; pron.; fém. plur.> indiquent <indiquer; verbe; ind. prés.; 3e pers. plur.> généralement <généralement; adv.>, en <en; prép.> premier <premier; adj.; masc. sing.> lieu <lieu; nom; masc. sing.>, la <le; dét.; fém. sing.> catégorie <catégorie; nom; fém. sing.> grammaticale <grammatical; adj.; fém. sing.> de <de; prép.> tous <tout; adj. masc. plur.> les <le; dét.; masc.; plur.>

```
térêt de l'exploration vers une   maladie   locale. Cette radiographie des sinu
le d'éviter l'évolution vers la   maladie   postphlébitique. La strycardiomyopa
le d'éviter l'évolution vers la   maladie   postphlébitique. La streptokinase e
L'adrénoleucodystrophie est une   maladie   récessive liée au sexe qui débute e
omalie et/ou trouble du rythme;   maladie   rythmique de l'oreillette. Dans la
ellement l'adulte alors que les   maladies  dysmyélinisantes se rencontrent hab
ellement l'adulte alors que les   maladies  dysmyélinisantes se rencontrent hab
  Les leucodystrophies sont des   maladies  métaboliques d'origine enzymatique
. Les leucodystrophies sont des   maladies  métaboliques d'origine enzymatique
t mitral, insuffisance mitrale,   maladies  mitrales, prolapsus valvulaire mitr
criniennes, il faut évoquer les   maladies  osseuses constitutionnelles dont la
```

Le corpus : extraction de contextes à partir d'étiquettes morphosyntaxiques (2)

WinBrill

Mot	Étiquette	Explication de l'étiquette
-	/-	ponctuation – tiret
rubéole	/SBC:sg	nom commun singulier
congénitale	/ADJ:sg	adjectif singulier
:	/ :	ponctuation – deux points
les	/DTN:pl	déterminant pluriel
signes	/SBC:pl	nom commun pluriel
cliniques	/SBC:pl	nom commun pluriel
d'	/SBC:sg	préposition
appel	/SBC:sg	nom commun singulier

TnT

Mot	Étiquette	Explication de l'étiquette
in	II	general preposition
atherosclerosis	NN	common noun
,	YC	punctuation, comma
deposits	NN	common noun
of	IO	of
plaque	JJ	general adjective
build	VV	verb, base form, present participle catenative, past participle
up	RP	particle
along	II	general preposition

Recherche et extraction de contextes (1) : isoler des contextes spécifiques

1. contextes définitoires

Elle affecte la rétine -- le tissu nerveux au fond de l'oeil qui transmet les messages visuels au cerveau (red_diab)

a green substance in leaves called chlorophyll. (contexte cité dans Pearson 1998)

2. contextes contenant des termes liés conceptuellement

Un disque virtuel est une partie de la mémoire configurée comme une disquette. ZMSPC2

Compost contains nutrients, nitrogen, potassium and phosphorous (contexte cité dans Meyer 2001).

Sulphur dioxide is a gas given off by some fuels (contexte cité dans Pearson 1998).

Recherche et extraction de contextes (2) : isoler des contextes spécifiques avec des marqueurs

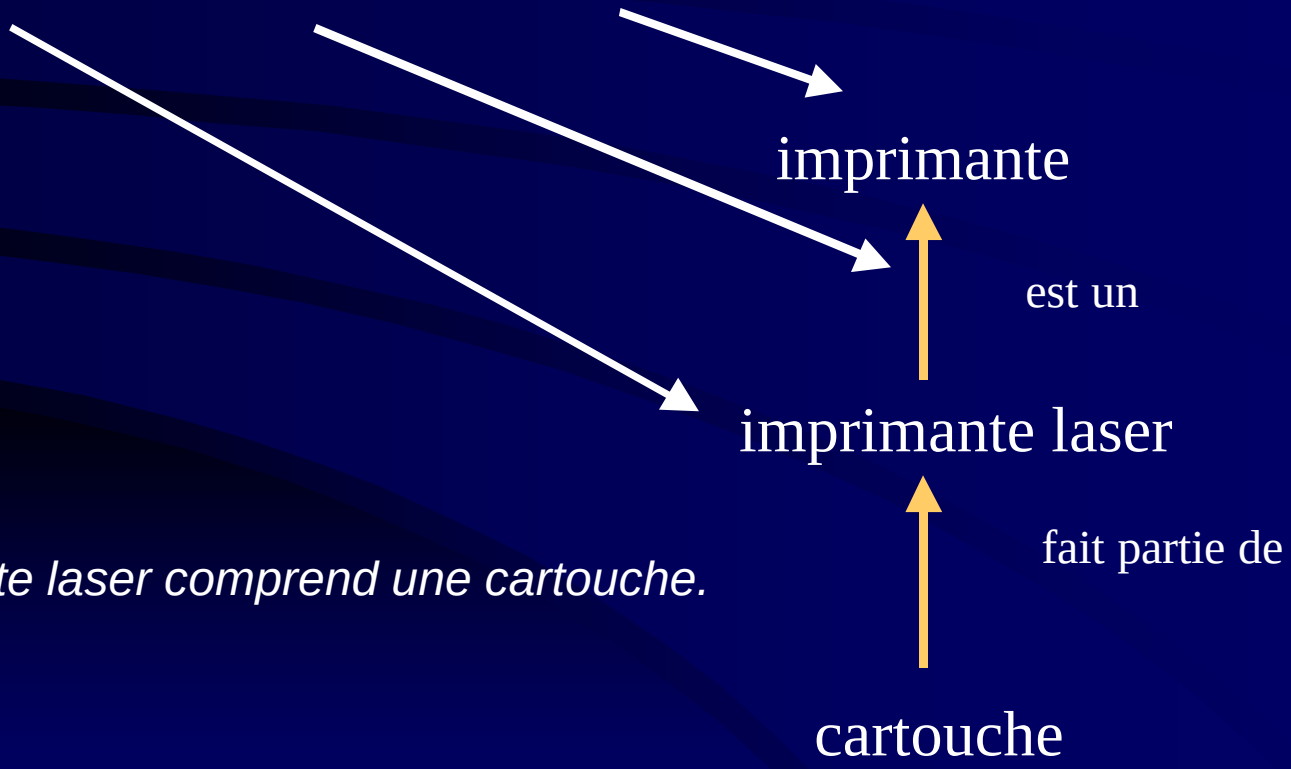
Terme + marqueur linguistique

Anglais	Français
is called, known as, also called	est appelé
is defined as	se définit de la manière suivante
consists of, contains, is part of	fait partie de, est constitué de, contient
is a, is a type of , is a form of	est un, est un type de
e.g., i.e., :, (*)	ex., c'est-à-dire, :, (*)

Exploitation de l'information contextuelle : détection automatique de liens conceptuels

Utilisation des marqueurs des liens conceptuels entre termes.

L'imprimante laser est un type d'imprimante



L'imprimante laser comprend une cartouche.

Dépistage de termes

Comment procéder pour isoler les termes à partir d'un corpus spécialisé ?

Forme des termes ?
- *cardiovasculaire*
- Termes complexes : *droit constitutionnel, mémoire programmable*

Catégorie grammaticale ? → Termes sont généralement des noms ou des syntagmes nominaux

Fréquence ? → Indices qui révèlent l'importance d'un mot d'une suite de mots dans un corpus ou l'affinité de deux mots

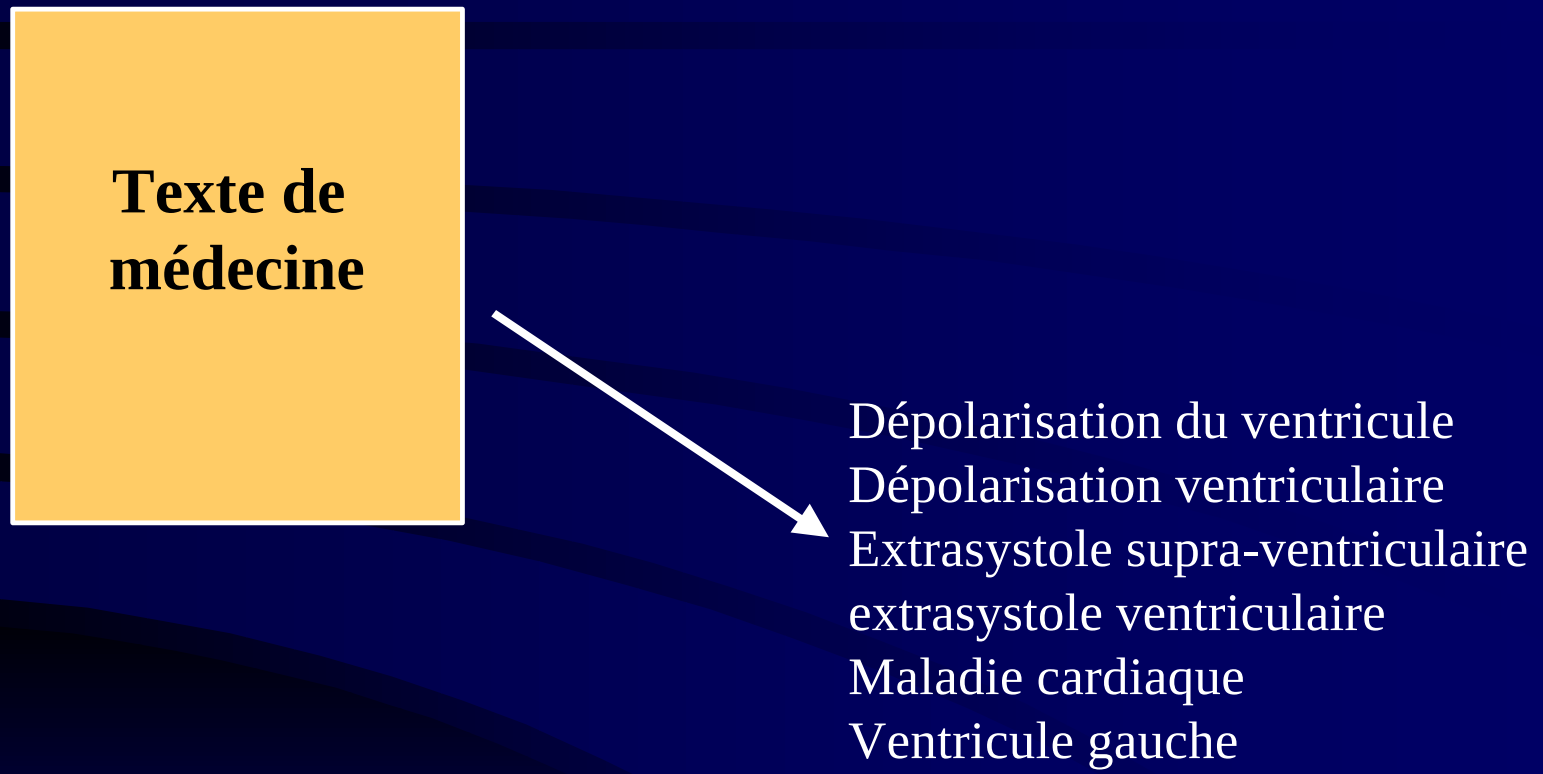
Éléments contextuels ? → Des indices dans la phrase : *est un, se définit comme*, etc.

Fréquence décroissante

Texte portant sur le droit constitutionnel
(suppression des mots grammaticaux)

droit	44
charte	15
liberté	14
état	13
personne	12
loi	12
homme	9
disposition	7
organe	7
Convention	6

Extraction terminologique (1)



Extraction terminologique (2) : 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 l
nom + (nom + adjectif), r
système de ges

Anglais

adjectif + nom : *volatile memory*

nom + nom : *laser printer*

*nom + préposition + nom : *extraction of terminology*

toute combinaison plus longue des combinaisons ci-dessus.

Extraction terminologique (3) : approche linguistique

a. Recherche de combinaisons de mots correspondant à des patrons

ex. nom + adjectif ; nom + prép. + nom (mettre dans la liste)
L'ordinateur portab
bureau.

Anglais

adjectif + nom ou nom + nom

This is the case in machine translation, as well as in computer-aided translation, automatic abstracting, text generation, etc.

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).

Extraction terminologique (4) : 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. un texte étiqueté

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

Extraction terminologique (5) : 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

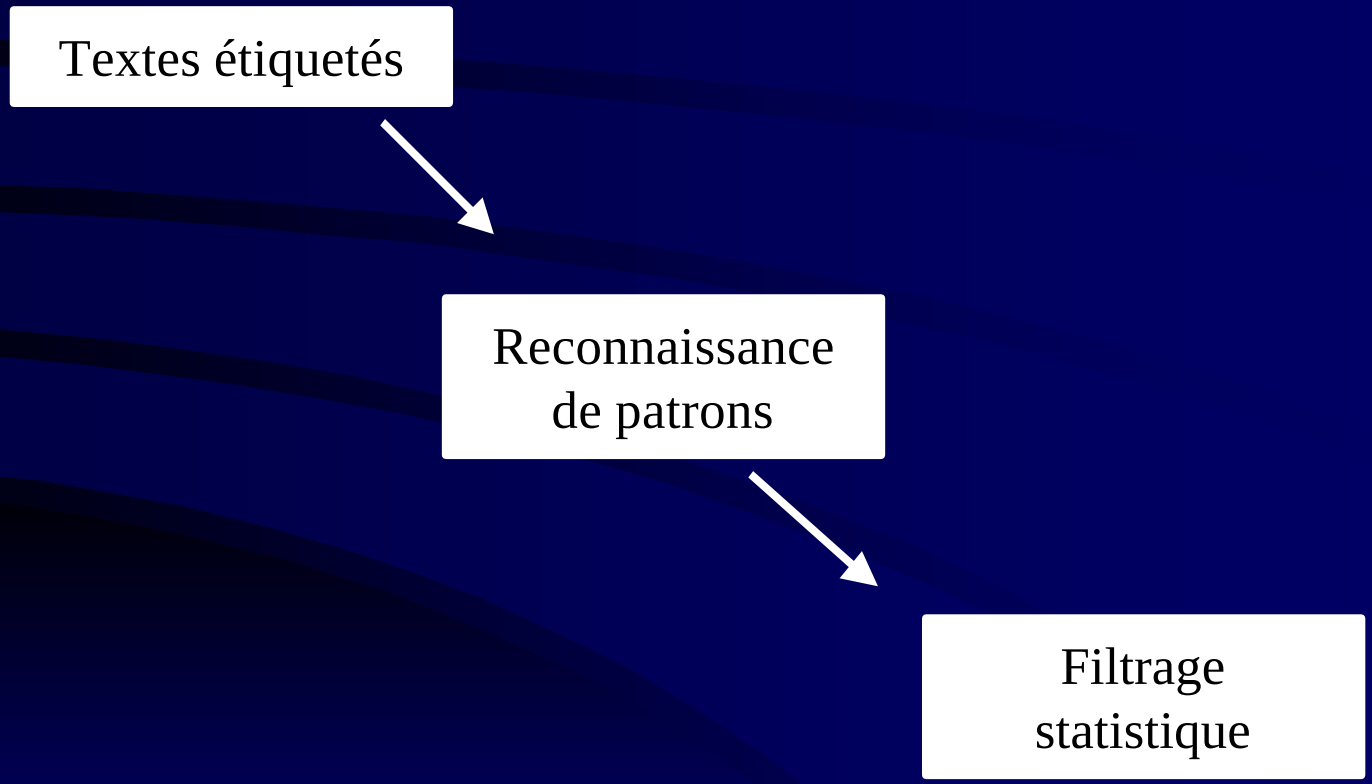
Si des mots simples apparaissent plus de n fois ensemble dans un texte, la suite est significative.

Exemple 2 : Information mutuelle

Si un mot X apparaît plus fréquemment dans l'entourage d 'un mot Y qu'ailleurs dans le texte, alors X et Y forment une combinaison significative.

Extraction terminologique (6) : approches hybrides

Un exemple : Daille (1994)



Extraction terminologique (7) : enrichissements récents

1. Extraction de termes simples : spécificités lexicales et corpus de référence

2. Regroupements de termes

- variantes terminologiques : cellule sanguine, cellule du sang; basal area, area of basal; active site, active secondary site; blood plasma, blood serum and plasma (Jacquemin 2001)
- familles formelles : logiciel d'application, logiciel comptable, logiciel de navigation, application commerciale, configuration du logiciel, installation de l'application

3. Extraction bilingue (alignement de traductions à partir de textes alignés, recherche d'équivalents à partir de textes comparables)

Détection de liens syntagmatiques

on décide de porter une accusation différente, mais pour la même affaire; ou encore lorsque, l'accusé ayant été libéré à la suite de son enquête préliminaire, (Morel)

Se référant au droit commun, il rappelle que dans une mise en accusation, on incorpore les faits reprochés et le droit spécifiquement enfreint. (Hebert 1984).

Il en va de même lorsqu'une nouvelle accusation est portée après que l'avocat de la poursuite eut ordonné d'arrêter les procédures (Morel)

comme dans le cas d'une fouille rectale et même s'il s'agit d'une accusation grave, (Chevre)

Et la gravité de l'accusation peut avoir cet effet aussi (Chevre)

Extraction de groupes syntagmatiques

1. Collocations : configurer un logiciel, installation d' un programme, install a program, configuration of a printer, etc.

2. Syntagmes verbaux : disséquer le plateau rocheux en chevrons, observer une charge excessive en trouble (Bourigault et Fabre 2000); latch into, beef up, quirrel away (Blaheta & Johnson 2001).

a) approche linguistique : recherche de patrons (verbe + nom, verbe + préposition)

b) approche statistique : calculs de combinaisons de mots (mais qui tiennent compte d'éléments variables ou insérés)

c) approches hybrides

Construction (automatique) de classes

Si vous obtenez un message d'erreur lorsque vous lancez Windows ...

De nombreux jeux DOS peuvent être lancés directement depuis Windows ...

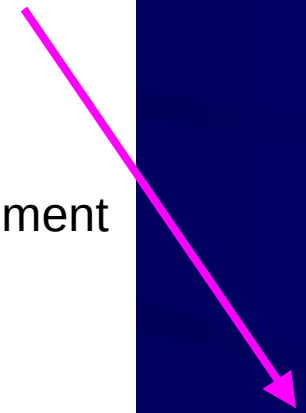
Lorsque vous lancez un programme, il est normalement utile de préciser son extension.

Lorsque vous lancez un logiciel en frappant la commande adéquate à partir du clavier ...

... un double clic sur un nom dans le gestionnaire de fichiers lance l'application correspondante ...

... lorsqu'on lance un programme, il ne va pas directement en mémoire ...

Lorsque WordPerfect est lancé, des indications sont affichées ...



lancer : application

jeux DOS

logiciel

programme

Windows

WordPerfect

Impacts sur le travail du terminographe (1)

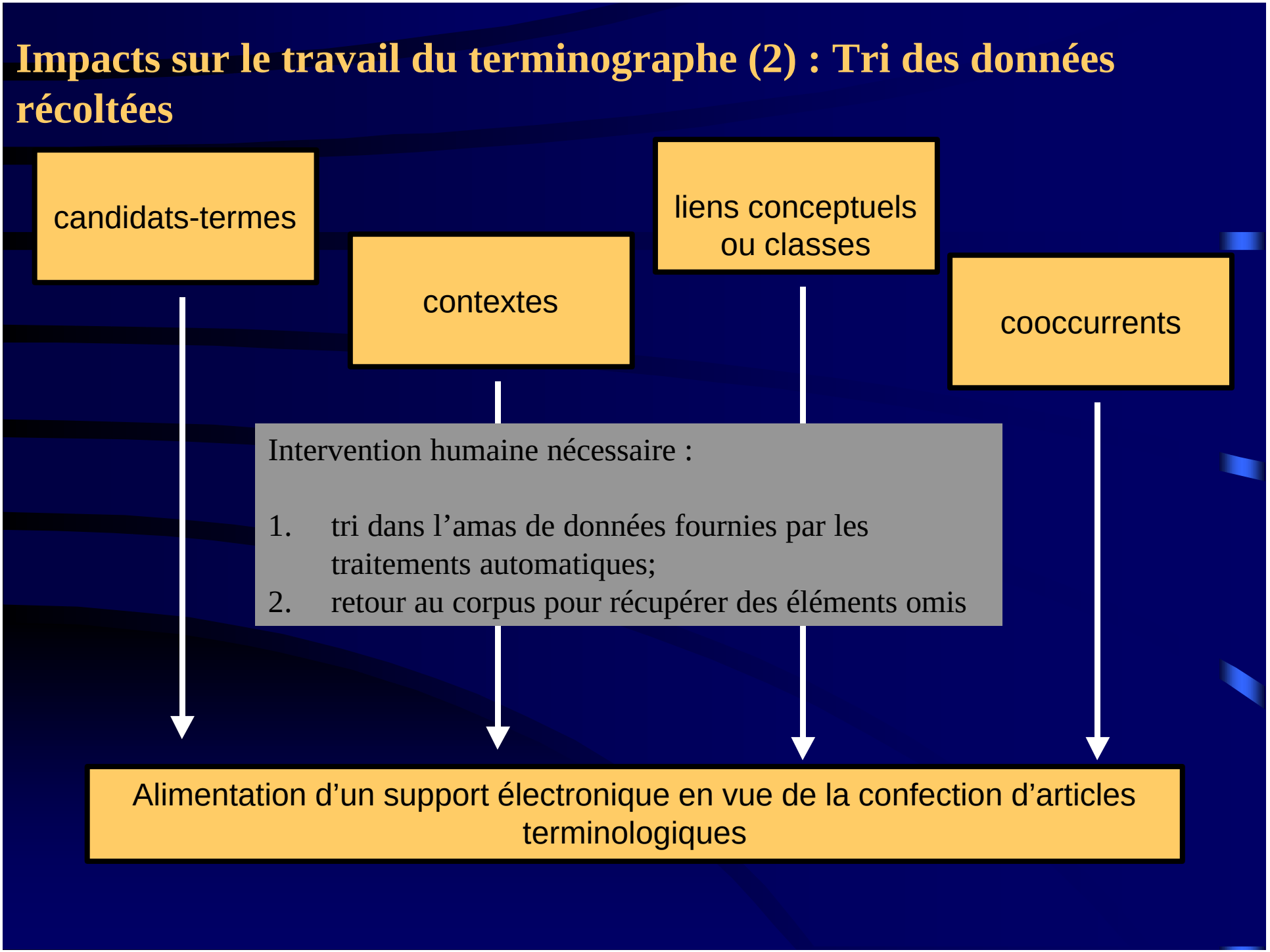
Facilite le travail du terminographe ?

Les logiciels d'extraction de termes traitent de grands volumes de textes spécialisés rapidement;

Les logiciels qui proposent d'autres types d'extraction (termes reliés conceptuellement, contextes ciblés, cooccurrents, etc.) sont utiles dans un contexte où les terminographes se penchent de plus en plus sur de nouveaux types de données;

Les corpus spécialisés sont extrêmement riches en information et l'extraction ne peut être faite uniquement par le terminographe.

Impacts sur le travail du terminographe (2) : Tri des données récoltées

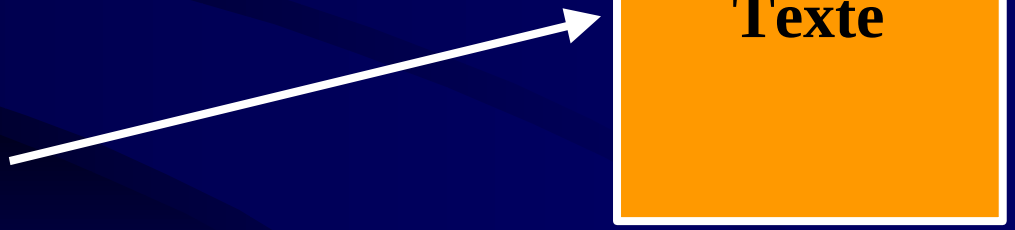


Impacts méthodologiques sur le travail du terminographe (3)

En recherche terminologique classique, l'extraction de contextes et de termes se font simultanément.

En recherche terminologique automatisée, l'extraction de contextes succède à l'extraction de termes.

computational linguistics
machine translation
specialized language
suitable equivalence
technical text
term recognition
terminological acquisition



Le retour au contexte est souvent indispensable pour déterminer la nature terminologique d'une unité.

Impacts méthodologiques sur le travail du terminographe (4) : acquisition de connaissances

L’acquisition de connaissances doit être faite séparément de l’extraction de termes (auparavant, les connaissances étaient acquises au moment du repérage);

Paradoxe : il faut très bien connaître le domaine traité pour juger de la pertinence d’un terme dans une liste proposée par un logiciel d’extraction (et a fortiori pour juger la nature d’une relation conceptuelle proposée ou d’un groupe syntagmatique).

Impacts méthodologiques sur le travail du terminographe (5) : gain de temps ?

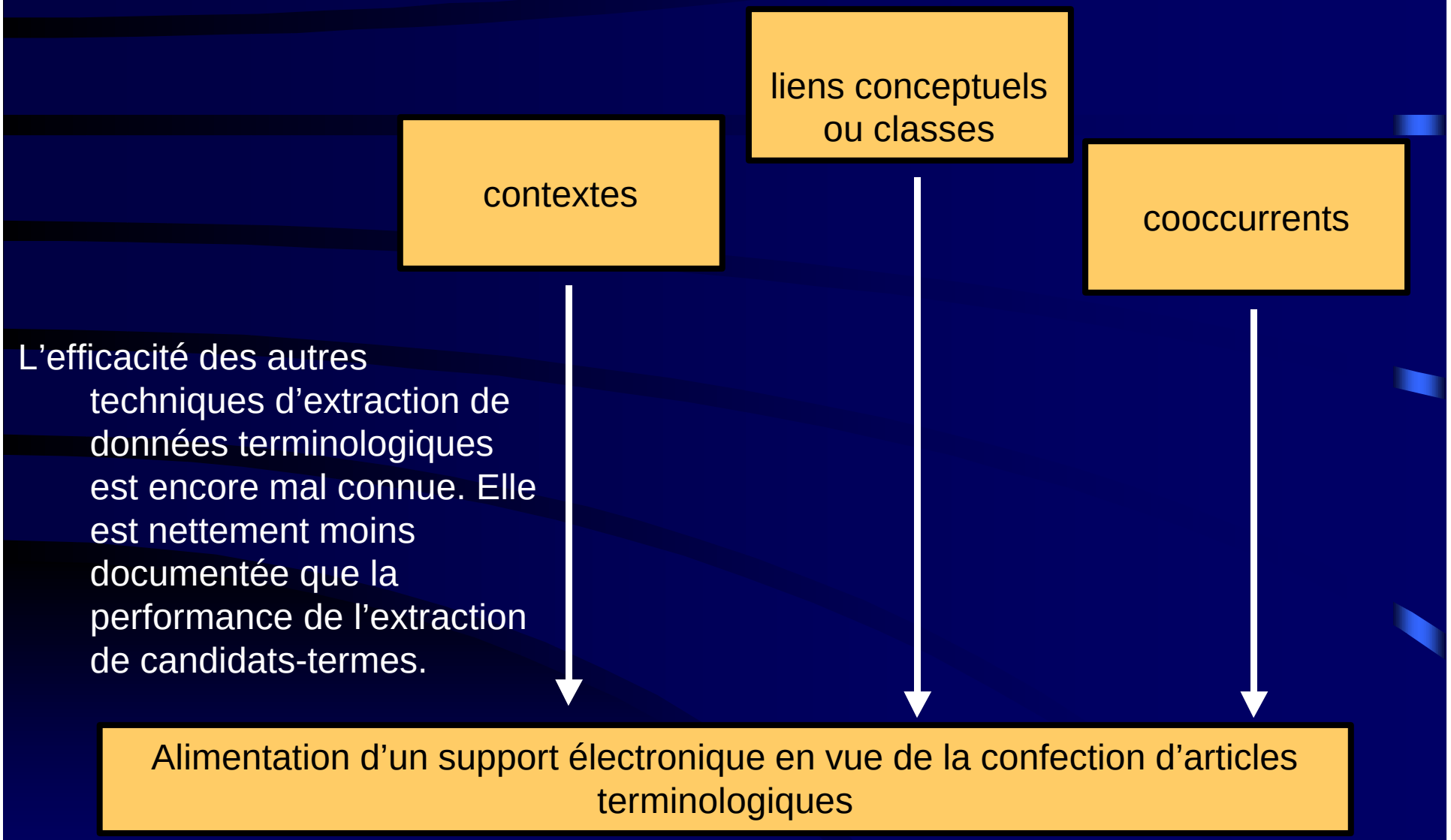
candidats-termes

On dispose de données sur la performance des logiciels d'extraction de termes. Toutefois, ces données sont fondées sur un examen des listes produites par ces extracteurs. On ne dispose pas de renseignements sur le rapport automatique-humain.

Contrairement aux concordanciers qui sont immédiatement appréciés, on connaît mal l'accueil réservé aux extracteurs de termes.

Alimentation d'un support électronique en vue de la confection d'articles terminologiques

Impacts méthodologiques sur le travail du terminographe (6) :



Conclusion

1. Toutes les données terminologiques ne peuvent être extraites d'un corpus (ex. données chronologiques, géographiques ou d'usage) à moins que le corpus n'air été préalablement caractérisé. L'expertise du terminographe est nécessaire à toutes les étapes. Le recours à des informateurs spécialistes est également indispensable.
2. Le recours à l'informatique, peu importe son intérêt, est incontournable en raison du nombre croissant de textes électroniques. S'il est impossible de l'éviter, il convient de mieux définir sa place dans les tâches terminologiques (réviser des méthodes de travail, par exemple).
3. Les outils d'extraction (et d'autres outils, notamment les outils de consignation des données) amènent le terminographe à envisager d'autres types de données terminologiques (amènent le terminographe à faire des coupes sélectives dans les corpus).
4. La formation en terminologie doit intégrer des notions de traitement automatique de langue minimale, de gestion de corpus et d'organisation des données dans des bases de données ou langages documentaires.