Expériences d'acquisition automatique de connaissances morphologiques par amorçage à partir d'un thésaurus

Pierre Zweigenbaum **Natalia Grabar**

DIAM

Service d'Informatique Médicale/DSI,

Assistance Publique – Hôpitaux de Paris

& Département de Biomathématiques, Université Paris 6

`{pz,ngr}@biomath.jussieu.fr`

`http://www.biomath.jussieu.fr/`

Plan

- Acquisition de connaissances morphologiques
- Matériel : SNOMED, CIM-10
- Méthodes:
  - À partir d'une liste de mots "à plat"
  - À partir de termes synonymes dans un thesaurus
    * Ajustement des suffixes des règles
    * Structuration des règles
- Évaluation des résultats
- Discussion et conclusion

Quelques définitions

Morphème unité minimale de sens

Proche d'un "concept"

Flexion variation des mots selon les règles de la grammaire

{*pied*, *pieds*}, {*généralisé*, *généralisée*}

Dérivation construction de mots : *base* + *affixe*

dentaire; {*abdomen*, *abdominal*}

Composition construction de mots ; composition savante

adénoacanthome achondrogenèse abêtalipoprotéinémie

Importance des ressources morphologiques

- Decomposition en morphèmes - decomposition en concepts
- Meilleure indexation potentielle des documents médicaux
- Aider à structurer la terminologie médicale
  relations morpholexicales → relations conceptuelles
- Amélioration potentielle des systèmes d'aide à la codification ou des serveurs de terminologie
- Traiter des mots inconnus en traitement automatique des langues / aider à construire un lexique pour le TAL :
  le lexique de la médecine n'est pas clos

Ressources morphologiques disponibles : un état des lieux contrasté

	Anglais	Allemand	Français	Russe
Flexion	++	++	++	??
Dérivation (général)	++	++	–	??
Dérivation (médical)	++	??	–	??
Composition (médical)	–	–	–	–

Il existe un besoin de construction de ressources morphologiques pour la dérivation et la composition.

Acquisition automatique de ressources morphologiques

Travaux antérieurs :

- à partir de corpus (Xu & Croft, 1998)
- à partir de thesaurus et corpus (Jacquemin, 1997)
- à partir de couples de mots reliés morphologiquement (Theron & Cloete, 1997)
- à partir d'une liste de mots à plat (Déjean, 1998)
- à partir d'une liste de mots étiquetés syntaxiquement (Hathout, 1999)

Ce travail :

- À partir des séries de termes synonymes d'un thesaurus

Objectifs

Acquisition automatique de connaissances morphologiques

- à partir d'un thesaurus (SNOMED)
  - liste de mots à plat, ou
  - séries de termes synonymes
- sans connaissances linguistiques a priori (traitement de chaînes de caractères)
- méthode indépendante de la langue
  Français (Anglais, Russe)

- Acquisition de connaissances morphologiques
- Matériel: SNOMED, CIM
- Méthodes:
  - Travail sur une liste de mots à plat
  - Amorçage sur des séries de synonymes d'un thesaurus
    - Ajustement des règles
    - Structuration des règles
- Évaluation des résulats
- Synthèse et conclusion

Matériel: SNOMED

Répertoire d'anatomopathologie français :

12 555 termes, 2 344 séries de synonymes

Code	Classe	Terme
D0-10430	01	pemphigoïde, SAI
D0-10430	02	pemphigus bénin, SAI
D6-50530	01	déficit en galactose épimérase
D6-50530	02	galactosémie type III
F-C6000	01	cellule immunitaire
F-C6000	02	immunocyte
F-D0650	01	fonction hémostatique
F-D0650	02	hémostase
M-02712	01	consistance anormalement dure
M-02712	02	durcissement
M-02712	05	durci

Matériel: Liste de mots “à plat”

Liste de 8 875 formes provenant de la SNOMED et de la CIM-10

a	
abaissé	. . .
abandon	éviscération
abcès	évitante
abdomen	évolutif
abdominal	évolution
abdominale	évoquant
abdominales	évoqué
abdominaux	événements
abdomino	éxentération
. . .	être
	îlots

- Acquisition de connaissances morphologiques
- Matériel: SNOMED, CIM
- Méthodes:
  - Travail sur une liste de mots à plat
  - Amorçage sur des séries de synonymes d'un thesaurus
    - Ajustement des règles
    - Structuration des règles
- Évaluation des résultats
- Synthèse et conclusion

Liste de mots à plat (candidats préfixes)

Effectué sur la liste de formes en français : 8 874 formes

Recherche d'occurrences de couples de mots contrastés :

couple $\{S, PS\}$ $\Rightarrow$ "*P-*" est un préfixe candidat

améloblastome	améloblastome
angiectasie	angiectasie
blastome	blastome
ectasie	ectasie
micronodulaire	micronodulaire
nodulaire	nodulaire
transépidermique	transépidermique
épidermique	épidermique

Liste de mots à plat (candidats préfixes)

Effectué sur la liste de formes en français : 8 874 formes

Recherche d'occurrences de couples de mots contrastés :

couple $\{S, PS\}$ $\Rightarrow$ *“P-”* est un préfixe candidat

améloblastome		améloblastome
angiectasie	“amélo-”	angiectasie
blastome	“angi-”	blastome
ectasie		ectasie
micronodulaire		micronodulaire
nodulaire	“micro-”	nodulaire
transépidermique		transépidermique
épidermique	“trans-”	épidermique

Liste de mots à plat (candidats suffixes)

Effectué sur la liste de formes en français : 8 874 formes

Recherche d'occurrences de couples de mots contrastés :

couple $\{P, PS\}$ $\Rightarrow$ "-S" est un suffixe candidat

acido	acido
acidocétose	acidocétose
actuelle	actuelle
actuellement	actuellement
adéno	adéno
adénolymphome	adénolymphome
broncho	broncho
bronchogénique	bronchogénique
bulle	bulle
bulleux	bulleux

Liste de mots à plat (candidats suffixes)

Effectué sur la liste de formes en français : 8 874 formes

Recherche d’occurrences de couples de mots contrastés :

couple $\{P, PS\}$ $\Rightarrow$ “-S” est un suffixe candidat

acido		acido
	“-cétose”	
acidocétose		acidocétose
actuelle		actuelle
	“-ment”	
actuellement		actuellement
adéno		adéno
	“-lymphome”	
adénolymphome		adénolymphome
broncho		broncho
	“-génique”	
bronchogénique		bronchogénique
bulle		bulle
	“-ux”	
bulleux		bulleux

- Acquisition de connaissances morphologiques
- Matériel: SNOMED, CIM
- Méthodes:
  - Travail sur une liste de mots à plat
  - Amorçage sur des séries de synonymes d'un thesaurus
    - Ajustement des règles
    - Structuration des règles
- Évaluation des résultats
- Synthèse et conclusion

Séries de synonymes

Méthode en deux étapes :

1. Amorçage : apprentissage de règles morphologiques
2. Expansion : induction de règles et application à des données plus larges

Synonymes: Amorçage

Hypothèse

Deux mots sont reliés morphologiquement s'ils partagent un morphème commun.

Heuristiques

- ils partagent une chaîne de caractères initiale commune, et
- ils apparaissent dans un contexte sémantiquement contraint

$\rightarrow$ alignement de mots (formes) dans des séries de synonymes

Alignement et segmentation de mots

Code	Classe	Terme
D0-10430	01	pemphigoïde, SAI
D0-10430	02	pemphigus bénin, SAI
D6-50530	01	déficit en galactose épimérase
D6-50530	02	galactosémie type III
F-C6000	01	cellule immunitaire
F-C6000	02	immunocyte
F-D0650	01	fonction hémostatique
F-D0650	02	hémostase
M-02712	01	consistance anormalement dure
M-02712	02	durcissement
M-02712	02	durci
M-35300	01	embole
M-35300	02	embolie
M-35300	05	embolique

Alignement et segmentation de mots

Code	Classe	Terme
D0-10430	01	*pemphigoïde*, SAI
D0-10430	02	*pemphigus* bénin, SAI
D6-50530	01	déficit en *galactose* épimérase
D6-50530	02	*galactosémie* type III
F-C6000	01	cellule *immunitaire*
F-C6000	02	*immunocyte*
F-D0650	01	fonction *hémostatique*
F-D0650	02	*hémostase*
M-02712	01	consistance anormalement *dure*
M-02712	02	*durcissement*
M-02712	02	*durci*
M-35300	01	*embole*
M-35300	02	*embolie*
M-35300	05	*embolique*

Création de familles morphologiques

Regroupement sur la chaîne de caractères initiale commune

larynx/laryngé *larynx/laryngo* *laryngé/laryngo* *larynx/laryngée* *laryngo/laryngopharynx*	$\longrightarrow$	*larynx* *laryngé* *laryngée* *laryngo*	$\longrightarrow$	*larynx* *laryngé* *laryngée* *laryngo* *laryngopharynx*

et sur un mot commun

Séries de synonymes

Méthode en deux étapes :

1. Amorçage : apprentissage de règles morphologiques
2. Expansion : induction de règles et application à des données plus larges

Synonymes: Approche inductive

Concepts

Règles morphologiques

induction

catégorisation

Exemples

couples de mots
reliés morphologiquement

hémostase, hémostatique
larynx, laryngé
amyloïdose, amyloïde

Population de test

Couples de mots dans LxL
(L = liste de mots)

...	...
glie	glie
glio	glio
...	...
gliome	gliome
...	...
gliose	gliose
glissant	glissant
...	...

Couples de mots (appris par alignement)

détecter de nouveaux couples de mots reliés morphologiquement

Induction de règles morphologiques

Couple de mots		Règle
larynx/*laryngé*	$\longrightarrow$	*x/gé*
larynx/*laryngo*	$\longrightarrow$	*x/go*
laryngé/*laryngo*	$\longrightarrow$	*é/o*
larynx/*laryngée*	$\longrightarrow$	*x/gée*
laryngo/*laryngopharynx*	$\longrightarrow$	ϵ */pharynx*

Expansion : Application de règles morphologiques

Les règles morphologiques permettent de relier des formes de la liste de référence :

Les couples de mots initiaux y figurent...

Expansion : Application de règles morphologiques

Les règles morphologiques permettent de relier des formes de la liste de référence :

... et de nouveaux couples complètent les familles morphologiques.

- Acquisition de connaissances morphologiques
- Matériel: SNOMED, CIM
- Méthodes:
  - Travail sur une liste de mots à plat
  - Amorçage sur des séries de synonymes d’un thesaurus
    - Ajustement des règles
    - Structuration des règles
- Évaluation des résultats
- Synthèse et conclusion

Ajustement des règles

Rendre les chaînes finales (suffixes) linguistiquement plus pertinentes pour se rapprocher de morphèmes

Règle initiale
e|que
se|x
le|ux
f|ve
me|sarcome
f|on
on|ve
|ux
e|ques
me|ïde
blastome|me
|use
f|ves
ateux|e

Ajustement des règles

Rendre les chaînes finales (suffixes) linguistiquement plus pertinentes pour se rapprocher de morphèmes

- Pendant l'expansion : on étend les chaînes finales au maximum vers la gauche (produit le même nombre de couples de mots)

Règle initiale	Règle ajustée
e|que	*ie|ique*
se|x	*euse|eux*
le|ux	*ale|aux*
f|ve	*if|ive*
me|sarcome	*ome|osarcome*
f|on	*if|ion*
on|ve	*ion|ive*
|ux	*e|eux*
e|ques	*ie|iques*
me|ïde	*ome|oïde*
blastome|me	*oblastome|ome*
|use	*e|euse*
f|ves	*if|ives*
ateux|e	*omateux|ome*

- Acquisition de connaissances morphologiques
- Matériel: SNOMED, CIM
- Méthodes:
  - Travail sur une liste de mots à plat
  - Amorçage sur des séries de synonymes d'un thesaurus
    * Ajustement des règles
    * Structuration des règles
- Évaluation des résultats
- Synthèse et conclusion

Structuration des règles

- Ordre de parcours : réduction de la taille des mots

  embolique $\xrightarrow{que|e}$ embolie

- Réduire la redondance des règles

  embolie $\xrightarrow{ie|e}$ embole

  embolique $\xrightarrow{que|e}$ embolie

  embolique $\xrightarrow{ique|e}$ embole

  En cas de réduction multiple, conserver la réduction de coût minimal

  Coût d'une règle = somme des longueurs de ses deux préfixes

- Segmentation résultante

  embolique $\longrightarrow$ embolie + *"-que"*

  embolie $\longrightarrow$ embole + *"-ie"*

- Acquisition de connaissances morphologiques
- Matériel: SNOMED, CIM
- Méthodes:
  - Travail sur une liste de mots à plat
  - Amorçage sur des séries de synonymes d'un thesaurus
    - Ajustement des règles
    - Structuration des règles
- Évaluation des résultats
- Synthèse et conclusion

Liste de mots à plat

Préfixes

	Total	$\geq$3 occ.	& longueur$\geq$2 char.
Trouvés	779	120	111
Corrects		107	106
Précision		89,2 %	95,5 %

“intra-” “in-” “péri-” “para-” “hyper-” “pré-” “hypo-” “trans-” “inter-” “fibro-” “ostéo-” “dé-” “neuro-” “micro-” “anti-” “poly-” “épi-” “sub-” “rétro-” “ré-” “pro-” “endo-” “dys-” “angio-” “myo-” “di-” “bi-” “pseudo-” “myélo-” “hém-” “extra-” “adéno-”

Séries de synonymes

Longueur minimale de la chaîne initiale commune = 3 caractères

Type de données	Amorçage
Nombre de termes	12,555
Séries de synonymes	2 344
Couples de mots	1 446
Couples de mots (différents)	1 087
Familles	755
Mots par famille	2,53

Séries de synonymes

Longueur minimale de la chaîne initiale commune = 3 caractères

Type de données	Amorçage	Expansion
Nombre de termes	12,555	
Séries de synonymes	2 344	
Liste de référence		8 875
Couples de mots	1 446	4 396
Couples de mots (différents)	1 087	4 396
Règles morphologiques	566	566
Familles	755	1 685
Mots par famille	2,53	3,13

Séries de synonymes

Longueur minimale de la chaîne initiale commune = 3 caractères

Type de données	Amorçage	Expansion	Précision
Nombre de termes	12,555		
Séries de synonymes	2 344		
Liste de référence		8 875	
Couples de mots	1 446	4 396	
Couples de mots (différents)	1 087	4 396	99%
Règles morphologiques	566	566	
Familles	755	1 685	95%
Mots par famille	2,53	3,13	

Erreurs

chrome	me/nique	chrome
chronique		chronique
crime	me/se	crime
crise		crise
entre	e/ée	entre
entrée		entrée
hyperplasie	plasie/kératose	hyperplasie
hyperkératose		hyperkératose
vers	ε/ion	vers
version		version

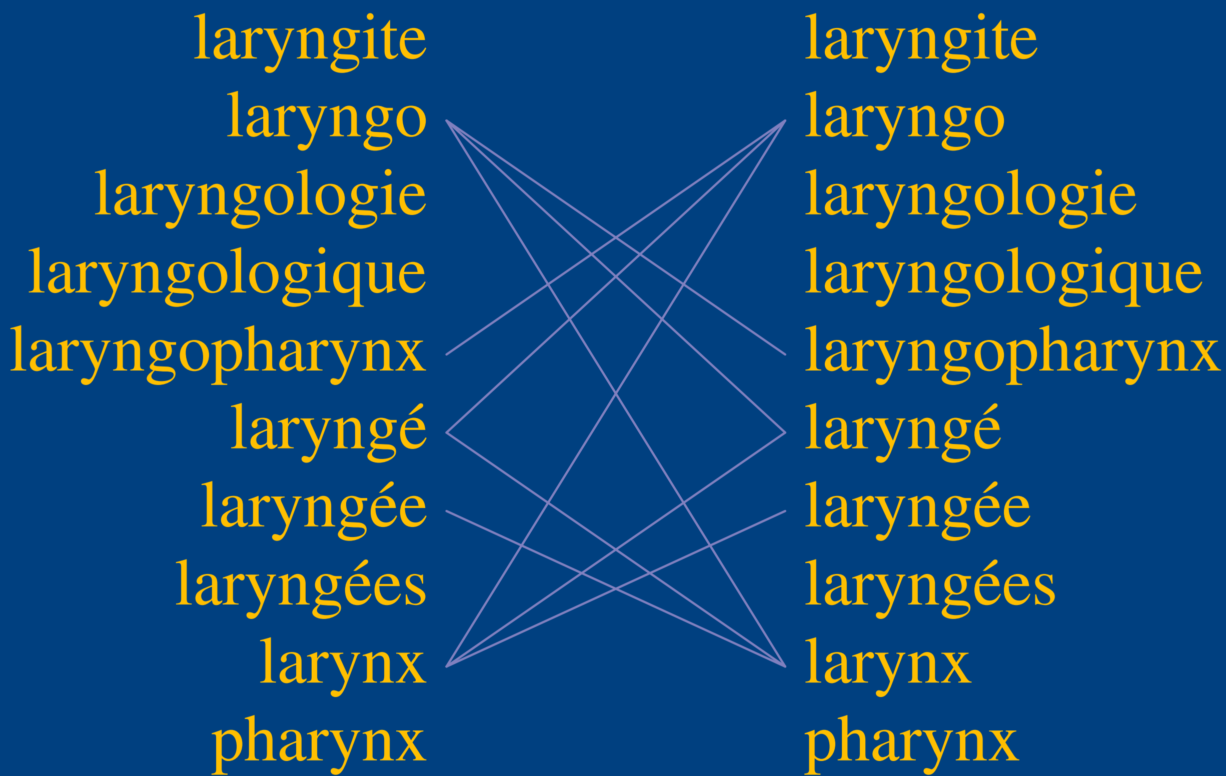
- Acquisition de connaissances morphologiques
- Matériel: SNOMED, CIM
- Méthodes:
  - Travail sur une liste de mots à plat
  - Amorçage sur des séries de synonymes d'un thesaurus
    * Ajustement des règles
    * Structuration des règles
- Évaluation des résultats
- Synthèse et conclusion

Synthèse

- Rétroacquisition de connaissances linguistiques fournies explicitement ou implicitement dans une terminologie
  $\rightarrow$ Excellente précision des couples de mots reliés morphologiquement
- Application de règles induites à la liste de mots
  $\rightarrow$ Augmentation du rappel
- Appliqué au français (ainsi qu'à l'anglais et au russe)
- L'étude contrastive de la liste de mots "à plat" fournit des segmentations et affixes complémentaires

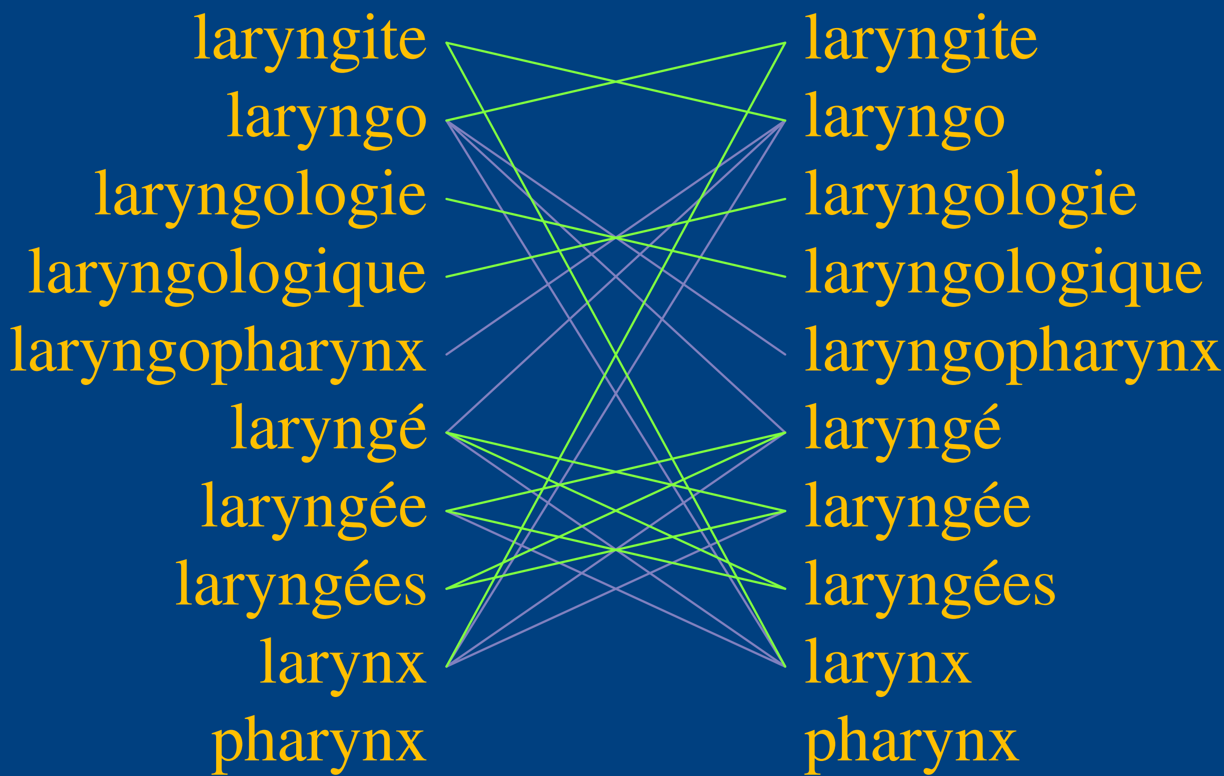
Augmentation du rappel

amorçage



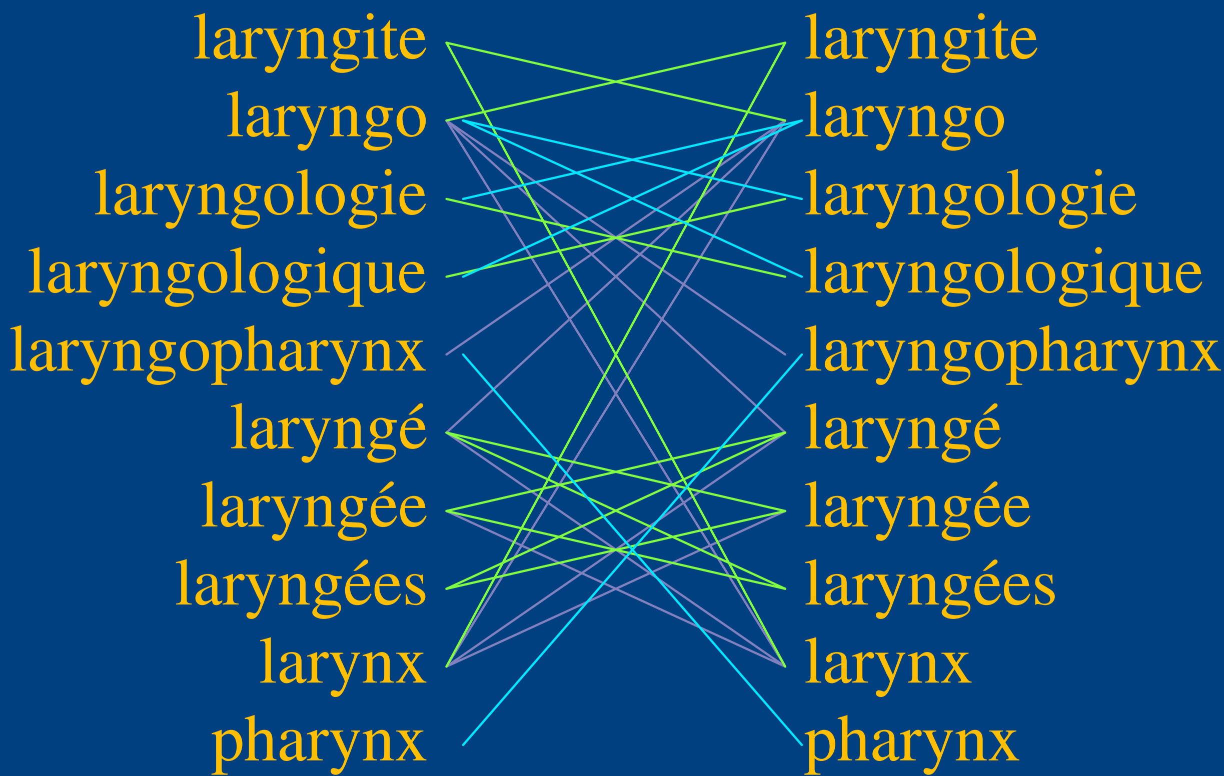
Augmentation du rappel

amorçage induction et application de règles



Augmentation du rappel

amorçage induction et application de règles liste de mots à plat



D'autres améliorations

- Seuil plus grand (taille de la chaîne initiale commune)
  → une meilleure précision
- A posteriori : ajustement de suffixes
  → segmentation plus correcte

Également testé (IMIA 1999) :

- Un étiquetage des mots ajoute des contraintes syntaxiques
  → meilleure précision
- Une lemmatisation élimine la flexion
  → règles plus propres (seulement dérivation et composition)

Perspectives

- Apprentissage inter-langue
  ← utiliser des versions parallèles de terminologies internationales
- Des règles plus adéquates
  ← un modèle morphologique plus élaboré (*e.g.*, morphologie à deux niveaux)
- Une structure sémantique interne pour les mots
  ← utilisation des axes et types sémantiques des terminologies
- Application à d'autres langues