TAL – Traitement Automatique des Langues

Natalia Grabar
natalia.grabar@univ-lille3.fr
http://natalia.grabar.free.fr

CNRS UMR8163 STL, Université Lille 3

Octobre 2017

TAL – Traitement Automatique des Langues

Objectifs du cours

- Découvrir le TAL
- Avoir une idée des tâches, méthodes, outils, applications
  - avoir une réflexion sur les méthodes
  - pouvoir utiliser les outils
- Focalisation sur certains points
- 4 jours : 8*3h de cours ou TD/TP (en alternance)
- 2 intervenants: Vincent Claveau, Natalia Grabar
- Contrôle de connaissances : projet noté
  - http://people.irisa.fr/Vincent.Claveau/cours/adt/
  - choix : lundi 16 octobre
  - rendu : mi-décembre

TAL – Traitement Automatique des Langues

Plan

1. Introduction au TAL
2. Ressources et connaissances terminologiques
   - accès aux informations terminologiques contenues dans les textes
     - termes
     - relations
     - manipulation de corpus textuels et d'outils de TAL spécialisés
3. Simplification de contenus spécialisés
   - diagnostic de la difficulté
   - acquisition de ressources

Les grands débuts de TAL

- Les années 1950 (guerre froide)
- Traduction automatique:
  - automatisation de la traduction d'une langue vers une autre
  - comprendre ce que disent les ennemis
- Environ $20 millions investis en 10 ans
- Test:
  - *The spirit is willing, but the flesh is weak*
  - $\Rightarrow$ Russe $\Rightarrow$ Anglais
  - *The whisky is strong, but the meat is rotten*

Dessous linguistiques

- Dictionnaire électronique
- Substitution de mots équivalents dans la langue cible transfert lexical
- Ordre syntaxique des mots
- Problématiques:
  - Ambiguïtés, polysémies, ...
  - Structures syntaxiques complexes
  - Relations sémantiques
  - Anaphores, ...

The "ALPAC report"

- En 1966, by the US National Academy of the Sciences
  Y. Bar-Hillel
  - Ni la bonne qualité ni l'automatisation complète ne peuvent être atteintes
  - L'automatisation complète n'est pas souhaitable :
    coûts potentiellement plus élevés qu'avec les traducteurs humains
  - *"MT is hopeless"*
  - Recommandation:
    - mettre plus d'effort dans la recherche en linguistique
    - qu'elle contribue ou non à la traduction automatique directement

$\Longrightarrow$ Début des travaux en TAL

Contributions

Un domaine interdisciplinaire:

- mathématiques:
  - logique
  - théorie des langages
  - probabilités
- informatique
  - algorithmique
  - génie logiciel
- linguistique
  - paramètres phonologiques
  - grammaire générative
  - syntaxe structurale
  - philosophie du langage

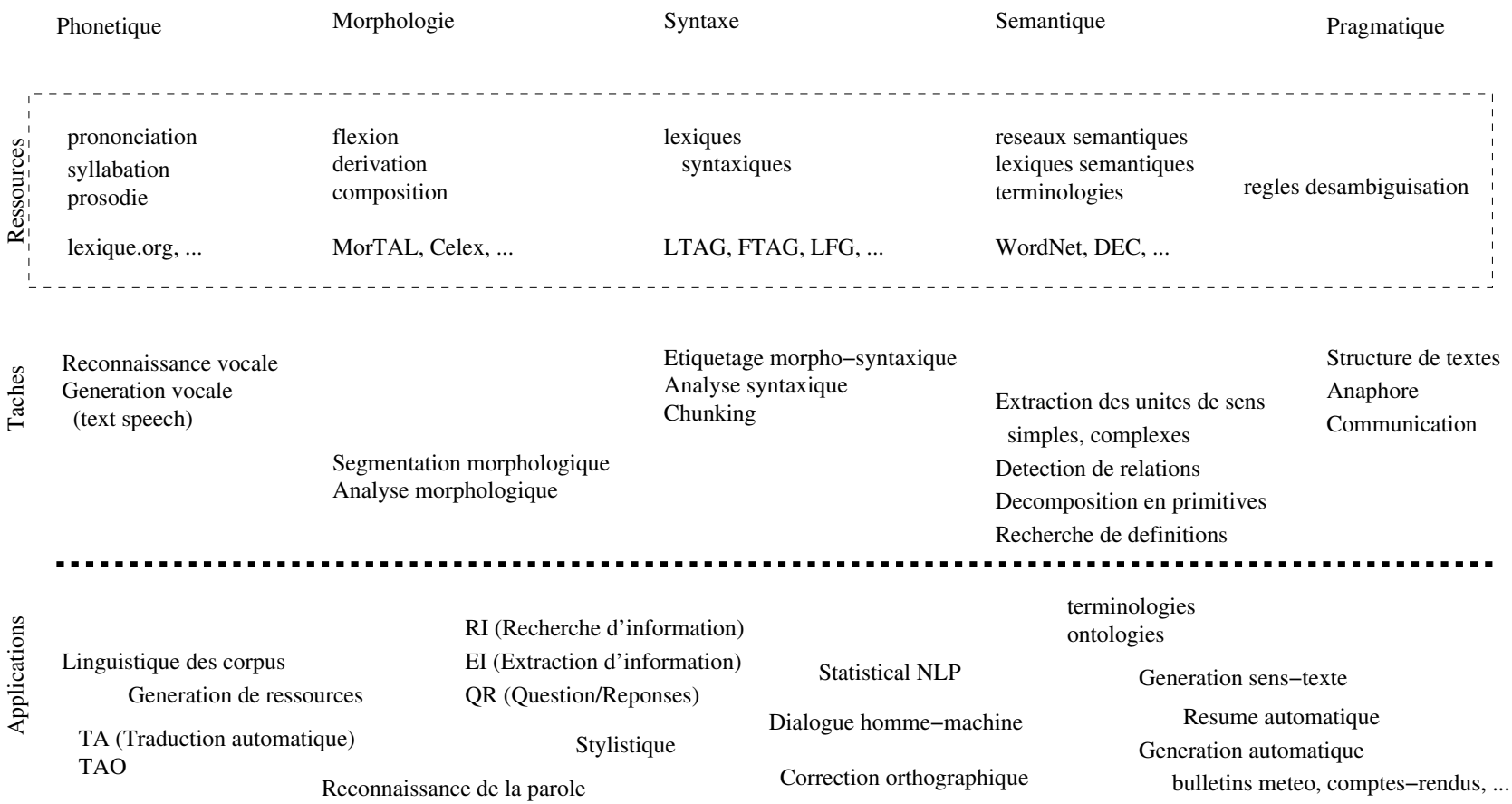
Répartitions

- Réparti dans les deux disciplines :
  - 1960 Linguistique informatique
    Focalisée sur les théories mathématiques, linguistiques
  - 1965 Traitement automatique des langues
    Focalisée sur les outils
- 1970 Natural Language Understanding (AI)
  approches cognitives
  - T Winograd, M Minski, J Allen, ...

50 ans plus tard

- Phonétique, phonologie, prosodie
- Morphologie
- Syntaxe
- Sémantique
- Pragmatique

50 ans plus tard



France

- ATALA: Association pour le traitement automatique des langues
- TAL: revue
- TALN, RECITAL, TALS, TALC, JEP, ...: conférences
- www.atala.org: site
- In: liste de diffusion
- Fillières de formations (LMD)
- Recrutements CNRS
- Besoins en entreprise
- Enjeux toujours réels

International

- ACL: Association for computational linguistics
- ACL, JNLE, ...: revues
- ACL, COLING, EACL, NNACL, LREC, ...: conférences
- www.aclweb.org: site
- linguist: liste
- De très nombreuses universtités (LMD)
- Besoins en entreprise
- Enjeux toujours réels

Plan

1. Introduction au TAL
2. Ressources et connaissances terminologiques
   - accès aux informations terminologiques contenues dans les textes
     - termes
     - relations
     - manipulation de corpus textuels et d'outils de TAL spécialisés
3. Simplification de contenus spécialisés
   - diagnostic de la difficulté
   - acquisition de ressources

Introduction

- Objet de la terminologie
- Utilité des terminologies
- Métiers
- Terminologie et corpus
- Documentation, stockage
- Recrutement de documents pour constituer un corpus

Qu'est ce qu'une/que la terminologie ?

Définition générale : ensemble des termes d'un domaine d'activité (médecine, aviation, électricité, etc.)

- Contenu, méthodes, science
- Langue générale *vs.* Langues de spécialité
- Domaine de spécialité
  - spécificité thématique
    objets décrits : un contenu, une sémantique
  - spécificité socio-culturelle
    pratiques réglées : manière de dire, de faire

Exemple d'utilisation d'une terminologie

Identification des informations pertinentes pour une tâche donnée :

- Recherche d'information: identification d'expressions specifiques (termes) de chaque document, ajout de relations entre eux, ou regroupement dans des classes

```
Comparative Analyses of Hairpin Substrate Recognition by Escherichia
coli and Bacillus subtilis Ribonuclease P Ribozymes.

Previously, we reported that the substrate shape recognition of the
Escherichia coli ribonuclease (RNase) P ribozyme depends on the
concentration of magnesium ion in vitro. We additionally examined the
Bacillus subtilis RNase P ribozyme and found that the B. subtilis enzyme
also required high magnesium ion, above 10 mM, for cleavage of a hairpin
substrate. The results of kinetic studies showed that the metal ion
concentration affected both the catalysis and the affinity of the
ribozymes toward a hairpin RNA substrate.
```

Exemple d'utilisation d'une terminologie

Identification des informations pertinentes pour une tâche donnée :

- Recherche d'information: identification d'expressions specifiques (termes) de chaque document, ajout de relations entre eux, ou regroupement dans des classes

Comparative Analyses of Hairpin Substrate Recognition by Escherichia coli and Bacillus subtilis Ribonuclease P Ribozymes.

Previously, we reported that the substrate shape recognition of the Escherichia coli ribonuclease (RNase) P ribozyme depends on the concentration of magnesium ion in vitro. We additionally examined the Bacillus subtilis RNase P ribozyme and found that the B. subtilis enzyme also required high magnesium ion, above 10 mM, for cleavage of a hairpin substrate. The results of kinetic studies showed that the metal ion concentration affected both the catalysis and the affinity of the ribozymes toward a hairpin RNA substrate.

Exemple d'utilisation d'une terminologie

Identification des informations pertinentes pour une tâche donnée :

- Recherche d'information: identification d'expressions specifiques (termes) de chaque document, ajout de relations entre eux, ou regroupement dans des classes

Comparative Analyses of Hairpin Substrate Recognition by Escherichia coli and Bacillus subtilis Ribonuclease P Ribozymes.

Previously, we reported that the substrate shape recognition of the Escherichia coli ribonuclease (RNase) P ribozyme depends on the concentration of magnesium ion in vitro. We additionally examined the Bacillus subtilis RNase P ribozyme and found that the B. subtilis enzyme also required high magnesium ion, above 10 mM, for cleavage of a hairpin substrate. The results of kinetic studies showed that the metal ion concentration affected both the catalysis and the affinity of the ribozymes toward a hairpin RNA substrate.

Exemple d'utilisation d'une terminologie

Identification des informations pertinentes pour une tâche donnée :

- Recherche d'information: identification d'expressions specifiques (termes) de chaque document, ajout de relations entre eux, ou regroupement dans des classes

```
Comparative Analyses of Hairpin Substrate Recognition by Escherichia
coli and Bacillus subtilis Ribonuclease P Ribozymes.

Previously, we reported that the substrate shape recognition of the
Escherichia coli ribonuclease (RNase) P ribozyme depends on the
concentration of magnesium ion in vitro. We additionally examined the
Bacillus subtilis RNase P ribozyme and found that the B. subtilis enzyme
also required high magnesium ion, above 10 mM, for cleavage of a hairpin
substrate. The results of kinetic studies showed that the metal ion
concentration affected both the catalysis and the affinity of the
ribozymes toward a hairpin RNA substrate.
```

Exemple d'utilisation d'une terminologie

Identification des informations pertinentes pour une tâche donnée :

- Recherche d'information: identification d'expressions specifiques (termes) de chaque document, ajout de relations entre eux, ou regroupement dans des classes

Comparative Analyses of Hairpin Substrate Recognition by Escherichia coli and Bacillus subtilis Ribonuclease P Ribozymes.

Previously, we reported that the substrate shape recognition of the Escherichia coli ribonuclease (RNase) P ribozyme depends on the concentration of magnesium ion in vitro. We additionally examined the Bacillus subtilis RNase P ribozyme and found that the B. subtilis enzyme also required high magnesium ion, above 10 mM, for cleavage of a hairpin substrate. The results of kinetic studies showed that the metal ion concentration affected both the catalysis and the affinity of the ribozymes toward a hairpin RNA substrate.

Exemple d'utilisation d'une terminologie

Identification des informations pertinentes pour une tâche donnée :

- Recherche d'information: identification d'expressions specifiques (termes) de chaque document, ajout de relations entre eux, ou regroupement dans des classes

```
Comparative Analyses of Hairpin Substrate Recognition by Escherichia
coli and Bacillus subtilis Ribonuclease P Ribozymes.

Previously, we reported that the substrate shape recognition of the
Escherichia coli ribonuclease (RNase) P ribozyme depends on the
concentration of magnesium ion in vitro. We additionally examined the
Bacillus subtilis RNase P ribozyme and found that the B. subtilis enzyme
also required high magnesium ion, above 10 mM, for cleavage of a hairpin
substrate. The results of kinetic studies showed that the metal ion
concentration affected both the catalysis and the affinity of the
ribozymes toward a hairpin RNA substrate.
```

Exemple d'utilisation d'une terminologie

Identification des informations pertinentes pour une tâche donnée :

- Recherche d'information: identification d'expressions specifiques (termes) de chaque document, ajout de relations entre eux, ou regroupement dans des classes

Comparative Analyses of Hairpin Substrate Recognition by Escherichia coli and Bacillus subtilis Ribonuclease P Ribozymes.

Previously, we reported that the substrate shape recognition of the Escherichia coli ribonuclease (RNase) P ribozyme depends on the concentration of magnesium ion in vitro. We additionally examined the Bacillus subtilis RNase P ribozyme and found that the B. subtilis enzyme also required high magnesium ion, above 10 mM, for cleavage of a hairpin substrate. The results of kinetic studies showed that the metal ion concentration affected both the catalysis and the affinity of the ribozymes toward a hairpin RNA substrate.

Exemple d'utilisation d'une terminologie

- Extraction d'information: remplir un formulaire, un enregistrement de base de données, décrivant un *événement*

```
Previously, we reported that the substrate shape recognition of the
Escherichia coli ribonuclease (RNase) P ribozyme depends on the
concentration of magnesium ion in vitro. We additionally examined the
Bacillus subtilis RNase P ribozyme and found that the B. subtilis enzyme
also required high magnesium ion, above 10 mM, for cleavage of a hairpin
substrate. The results of kinetic studies showed that the metal ion
concentration affected both the catalysis and the affinity of the
ribozymes toward a hairpin RNA substrate.
```

Exemple d'utilisation d'une terminologie

- Extraction d'information: remplir un formulaire, un enregistrement de base de données, décrivant un *événement*

```
Previously, we reported that the substrate shape recognition of the
Escherichia coli ribonuclease (RNase) P ribozyme depends on the
concentration of magnesium ion in vitro. We additionally examined the
Bacillus subtilis RNase P ribozyme and found that the B. subtilis enzyme
also required high magnesium ion, above 10 mM, for cleavage of a hairpin
substrate. The results of kinetic studies showed that the metal ion
concentration affected both the catalysis and the affinity of the
ribozymes toward a hairpin RNA substrate.
```

Exemple d'utilisation d'une terminologie

- Extraction d'information: remplir un formulaire, un enregistrement de base de données, décrivant un *événement*

```
Previously, we reported that the substrate shape recognition of the
Escherichia coli ribonuclease (RNase) P ribozyme depends on the
concentration of magnesium ion in vitro. We additionally examined the
Bacillus subtilis RNase P ribozyme and found that the B. subtilis enzyme
also required high magnesium ion, above 10 mM, for cleavage of a hairpin
substrate. The results of kinetic studies showed that the metal ion
concentration affected both the catalysis and the affinity of the
ribozymes toward a hairpin RNA substrate.
```

Eléments de comparaison

avec d'autres ressources

Terminologie *vs.* ...

- Lexiques, dictionnaires généraux
- Encyclopédies
- Ontologies

Terminologie *vs.* lexiques, etc.

Lexiques

- apprentissage, traduction
- mono ou multilingues
- entrée lexicale
- caractéristiques linguistiques

  phonétique, morphologie, syntaxe
- traitements automatiques

Terminologie *vs.* lexiques, etc.

Dictionnaires lexicographiques, Dictionnaires de langue

(le Robert, Larousse)

- monolingue
- langue générale
- entrée dictionnairique
- caractéristiques linguistiques
- définition
- exemples d'emploi
- désambiguïsation

Terminologie *vs.* lexiques, etc.

Encyclopédies

- fonction : didactique
- entrée
- définition
- pas de spécificité thématique
- noms communs, noms propres
- illustrations

Terminologie *vs.* lexiques, etc.

Terminologies

(fiche terminologique)

- domaine donné
- doit répondre à un besoin/une fonction : normalisation, description, transmission
- entrée : terme
- définition, exemple
- père, fils (relations hiérarchiques)
- d'autres relations (synonymie, etc.)
- identifiant, illustrations

Terminologies *vs.* ontologies

- Terminologie :
  - Ensemble de termes représentant un système de concepts pour un domaine particulier
  - Association d'informations linguistiques aux termes (entités linguistiques)
  - Pas d'organisation formelle des termes (d'où d'éventuels problèmes de cyclicités ou de redondances)
- Ontologie :
  - Description formelle des concepts et des relations pour un domaine particulier
  - Association de propriétés aux concepts
  - Destinée à des traitements automatiques (possibilité d'inférence et calcul formel)

Deux ressources complémentaires mais ATTENTION à la confusion
Une terminologie peut servir à peupler une ontologie

Utilité des terminologies

- Terminologie normative
- Terminologie descriptive
- Terminologie comme dépôt de connaissances

Terminologie normative

Diderot, Encyclopédie

Constat :

La langue des arts est très imparfaite pour deux causes :

- la disette des mots propres
- l'abondance des synonymes

Dans la langue des arts, un marteau, une tenaille, une auge, une pelle ont presque autant de dénominations qu'il y a d'arts

Plus récemment (début du $20^{ème}$ siècle), Ecole de Vienne, théorie Générale de la terminologie (Wüster 1981) :

- Universalité des notions
- Facilité la communication et la traduction

Terminologie normative

Situation la plus courante : dénominations différentes

- Variation régionale
  - Une même langue
    recommandations (Fr) *vs. ligne directrice* (Ca)
  - Langues différentes
    maladie de Weber-Christian (Fr) *vs.*
    maladie de Pfeifer-Weber-Christian (Ca)
    Pfeifer-Weber-Christian-Syndrom/Pfeifer-Weber-Christian-Krankheit (De)
- Stratégie commerciale
  marquer la différence entre les produits similaires mais proposés par des industriels différents
  airbag vs. coussin de sécurité vs. coussin gonflable
  pompe à chaleur vs. pompe thermique vs. thermopompe
  heat pump vs. thermal pump vs. thermopump

Terminologie normative

Situation la plus courante : dénominations différentes

- Locuteur
  - Idiolecte
    sténose de l'aorte vs. sténose aortique
  - Spécialisation, formation
    infarctus du myocarde vs. crise cardiaque
    rhagade vs. crevasse
- Diachronie

  oculiste (1503) *vs. ophtalmologue* (1840) *vs. ophtalmologiste*
  base latine, base grecque

Terminologie normative

Contextes d'application :

- entreprises, organismes internationaux, etc.
- traducteurs, rédacteurs
- documentation technique, officielle, etc.
- avoir le même référentiel d'expressions, de termes
- contrôle de la langue technique
- interopérabilité sémantique

$\rightarrow$ enregistrement de termes normalisés

$\Rightarrow$ *Langue idéale*

$\Rightarrow$ *Langage contrôlé*

Terminologie descriptive

- traitements automatiques
- garantir un accès aux documents

→ besoin d'enregistrer l'état actuel de la langue

- réagir aux développements technologiques
  cogénération d'électricité et chaleur

⇒ *Langue libre*

Terminologie comme dépôt de connaissances

- entreprises
- départs à la retraite
- changement du personnel
- transmission de la connaissance, formation du personnel

$\rightarrow$ besoin d'enregistrer l'état actuel de connaissances

$\Rightarrow$ *Experts*
$\Rightarrow$ *Bases de données*, *Bases de connaissances*
$\Rightarrow$ *Connaissances collaboratives*

Métiers et utilisateurs

- Étudiants
- Grand public
- Traducteurs, rédacteurs techniques
- Documentalistes
- Normalisateurs
- Lexicographes, encyclopédistes, pédagogues
- Terminologues
- Ingénieurs de la connaissance
- Intelligence artificielle
- TAListes

Stratégies de constitution

- Interviews avec des experts
- Ressources dictionnairiques et terminologiques existantes
- Exploration de corpus

Plan

1. Introduction
2. *Exemples de terminologies* (domaine médical)
3. Constitution de terminologie à partir de corpus
    1. Extraction de termes
    2. Extraction de relations

Exemples de terminologies

domaine médical

- MeSH, indexation et recherche d'information
- CIM, classification de causes de maladies
- SNOMED, encodage de dossiers patient
- MedDRA, effets indésirables des médicaments
- UMLS, union de terminologies médicales

MeSH, indexation et recherche d'information

MeSH, Medical Subject Headings

- *Indexation de documents et recherche d'information*

en : 313 372 concept, 737 164 labels, 16 types de relation, 879 884 relations

fr : 38 622 concepts, 94 366 labels, 4 types de relation, 9 896 relations

- Synonymes, liens hiérarchiques et d'association

`http://www.nlm.nih.gov/mesh/meshhome.html`

MeSH, indexation et recherche d'information

Extrait de la partie anglaise

A01	Body Regions	D001829
A01.835	Skin	D012867
A01.835.180	Dermis	D020405
A01.835.180.360	Hair Follicle	D018859
A01.835.180.800	Sebaceous Glands	D012627
A01.835.180.830	Sweat Glands	D013545
A01.835.180.830.040	Apocrine Glands	D001050
A01.835.180.830.280	Eccrine Glands	D004439
A01.835.250	Epidermis	D004817
A01.835.288	Animal Fur	D006197
A01.835.288.296	Eyebrows	D005138
A01.835.288.421	Eyelashes	D005140

MeSH, indexation et recherche d'information

Extrait de la partie française

A01	REGION	D001829
A01.835	PEAU	D012867
A01.835.180	DERME	D020405
A01.835.180.360	FOLLICULE PILEUX	D018859
A01.835.180.800	GLANDES SEBACEES	D012627
A01.835.180.830	GLANDES SUDORIPARES	D013545
A01.835.180.830.040	GLANDE APOCRINE	D001050
A01.835.180.830.280	GLANDE ECCRINE	D004439
A01.835.250	EPIDERME	D004817
A01.835.288	CHEVEU	D006197
A01.835.288.296	SOURCIL	D005138
A01.835.288.421	CIL	D005140

CIM, classification de causes de maladies

CIM, Classification Internationale des Maladies
(Classification statistique Internationale des Maladies et des problèmes de santé connexes)

- OMS (Organisation Mondiale de la Santé)
- *indexation des dossiers patient pour les besoins statistiques*

en : 12 318 classes, 13 505 libellés

fr : 10 800 classes, 9 412 libellés

- termes inclus (synonymes), liens hiérarchiques

CIM – ICD, classification de causes de maladies

Extrait de la partie anglaise

A04	Other bacterial intestinal infections
A04.0	Enteropathogenic Escherichia coli infection
A04.1	Enterotoxigenic Escherichia coli infection
A04.2	Enteroinvasive Escherichia coli infection
A04.3	Enterohaemorrhagic Escherichia coli infection
A04.4	Other intestinal Escherichia coli infections
A04.5	Campylobacter enteritis
A04.6	Enteritis due to Yersinia enterocolitica
A04.7	Enterocolitis due to Clostridium difficile
A04.8	Other specified bacterial intestinal infections
A04.9	Bacterial intestinal infection, unspecified

CIM, classification de causes de maladies

Extrait de la partie française

A04	Autres infections intestinales bactériennes
A040	Infection entéropathogène à Escherichia coli
A041	Infection entérotoxigène à Escherichia coli
A042	Infection entéro-invasive à Escherichia coli
A043	Infection entéro-hémorragique à Escherichia coli
A044	Autres infections intestinales à Escherichia coli
A045	Entérite à Campylobacter
A046	Entérite à Yersinia enterocolitica
A047	Entérocolite à Clostridium difficile
A048	Autres infections intestinales bactériennes précisées
A049	Infection intestinale bactérienne, sans précision

SNOMED, encodage de dossiers patient

Nomenclature SNOMED Int.

(NOmenclature Systématique des MÉDecines humaine et vétérinaire)

- *description des informations cliniques* (dossiers patient)

en : 112 661 concepts, 164 069 labels

fr : 9 098 concepts, 12 554 labels,
11 290 relations hiérarchiques en français

- synonymes, liens hiérarchiques, méronymiques et transversaux

SNOMED, encodage de dossiers patient

Terminologie multiaxiale

Les concepts sont organisés hiérarchiquement en onze axes sémantiques :

- Morphologie (M)
- Topographie (T)
- Fonction (F)
- Organismes vivants (L)
- Agents et activités physiques (A)
- Diagnostics (D)
- etc.

SNOMED, encodage de dossiers patient

Extrait de la partie anglaise

M-12000	01	Fracture, NOS
M-12000	05	Fractured
M-12010	01	Fracture, transverse
M-12020	01	Fracture, oblique
M-12200	01	Fracture, open, NOS
M-12300	01	Fracture, ununited, NOS
M-12400	01	Fracture, delayed union, NOS
M-12500	01	Fracture, healed, NOS
M-12500	02	Fracture, united, NOS
M-12520	01	Fracture, healed, fibrous union
M-12590	01	Pseudoarthrosis, NOS
M-12590	02	Nearthrosis
M-12590	02	Neoarthrosis

SNOMED, encodage de dossiers patient

Extrait de la partie française

M-12000	01	fracture, SAI
M-12000	05	fracturé
M-12010	01	fracture transverse
M-12020	01	fracture oblique
M-12200	01	fracture ouverte, SAI
M-12300	01	fracture non consolidée, SAI
M-12400	01	fracture avec retard de consolidation, SAI
M-12500	01	fracture consolidée, SAI
M-12500	02	fracture guérie, SAI
M-12520	01	fracture avec consolidation fibreuse
M-12590	01	pseudarthrose, SAI
M-12590	02	néarthrose

SNOMED, encodage de dossiers patient

Post-coordination

Un diagnostic D

- `est-un` atteinte morphologique M
- `localisé-dans` une partie du corps A

pneumonie

- `est-un` *inflammation*
- `localisé-dans` (*poumon*)

otite

- `est-un` *inflammation*
- `localisé-dans` (*oreille*)

apendicite
apendicectomie

MeDDRA, effets indésirables des médicaments

Medical Dictionary for Drug Regulatory Activities
Objectifs :

- Décrire les étapes du développement des médicaments
- Décrire les problèmes liés aux affaires réglementaires

Utilisé par :

- FDA (États-Unis), bases de données AERS, VAERS
- EMA (European Medicine Agency), base Eudrawatch
- PEM (Prescription Event Monitoring), Japon
- Standard d'échange d'information en Europe

MeDDRA, effets indésirables des médicaments

Medical Dictionary for Drug Regulatory Activities termes :

- Effets indésirables médicamenteux
- Indications
- Signes et symptômes
- Histoire familiale
- Examens de laboratoire
- Interventions chirurgicales

5 niveaux hiérarchiques

UMLS, union de terminologies médicales

UMLS

(Unified Medical Language System)

- *Intégration de plusieurs terminologies médicales*
- environ 20 langues
- plus de 158 terminologies
- Interopérabilité sémantique entre les terminologies

en : 2 381 083 concepts, 5 491 897 labels

fr : 82 170 concepts, 156 762 labels

- synonymes, liens hiérarchiques, environ 100 liens sémantiques

UMLS, union de terminologies médicales

Composants d'UMLS :

- Metathesaurus
  terminologies sources
- Semantic Network
  relations sémantiques entre termes
- SPECIALIST Lexicon
  lexique anglais (général, médical)
- MetamorphoSys
  outils pour faciliter l'utilisation d'UMLS
  exploration de tables, visualisations, appariement de termes, ...

Une terminologie unique ?

En fonction de :

- Besoins
- Applications
- Possibilités
- Outils
- Experts
- ...

Plan

1. Introduction
2. Exemples de terminologies (domaine médical)
3. *Constitution de terminologie à partir de corpus*
    1. Extraction de termes
    2. Extraction de relations

Terminologie et corpus

Stratégies de constitution

- Interviews avec des experts
- Ressources terminologiques existantes

$\Rightarrow$ Exploration de corpus (de spécialité)

Corpus de spécialité

- domaine de spécialité
  médecine, cogénération, télécommunications, etc.
- représente un langage de spécialité (sous-langage)
  pour un domaine donné
- a des particularités lexicales et grammaticales
- sert de base pour la production de :
  - terminologies
  - thesaurus, etc.
- structure interne :
  sous-corpus thématiques, origine, etc.

Terminologie de corpus

Fondement principal des approches visant à constituer des terminologies

- Pas d'*a priori* sur la langue
- L'information nécessaire se trouve dans le corpus
- Les sous-langages se caractérisent par :
  - un lexique limité (termes, synonymes)
  - schémas syntaxiques particuliers
- Disponibilité d'outils nécessaires : dépouillement et traitements du corpus

Exigences vis-à-vis d'un corpus

Le corpus doit satisfaire une triple exigence :

- pertinence par rapport au domaine
  textes représentatifs de ceux produits dans le domaine
- pertinence par rapport à la tâche
  textes représentatifs de ceux manipulés par l'application finale
- prise en compte des possibilités des outils de traitement automatique

Pertinence par rapport au domaine

- Caractérisation du domaine :
  mots clés, termes centraux du domaine, descripteurs
- Garantie de la centralité des documents :
  présence des mots clés dans ces documents
- Recensement de textes qui véhiculent la connaissance du domaine :
  - textes spécialisés
  - textes pour les non-spécialistes
  - textes de vulgarisation
    (à inclure suivant la tâche)

Comment évaluer la pertinence par rapport au domaine ?

Quantifier les mots clés (Salton, 1991) :

- fréquence absolue
  nombre de mots clés dans le document :
  F_d
- fréquence pondérée par la longueur du document
  rapport entre F_d *et le nombre d'occurrences* :
  $\frac{F_d}{N_{occ}}$
- pondération normalisée dans le document :
  $0.5 + 0.5\frac{F_d}{F_{max}}$
- pondération par rapport au corpus :
  $tf * idf = F_d * log(\frac{N_{doc}}{n})$

Comment évaluer la pertinence par rapport au domaine ?

Pondérer le document avec des méthodes vectorielles $\Longrightarrow$ représenter un document par les descripteurs qu'il contient :

- pondérer les descripteurs avec une ou plusieurs méthodes
- pondérer le document :
  - $sim_{Dom,Doc} = \frac{\sum_{i=0}^{n} P_i}{N_{doc}}$
  - $sim_{Dom,Doc} = \frac{\sum_{i=0}^{n} P_{ti}}{\sqrt{\sum_{i=0}^{n} P_{ti}^2 N_{td}}}$

  P_i poids d'un descripteur dans le document
  N_{td} nombre total de descriteurs dans le domaine

Pertinence par rapport à la tâche

Prendre en compte l'application finale :

- objectifs
  un phénomène linguistique, grammaire, style littéraire
- application
  recherche d'information, indexation, etc.
- portée
  interne, externe à une entreprise, nationale, etc.
- spécialisation
  corpus de langue générale, de langue de spécialité
- type de corpus
  contexte multilingue : corpus parallèle

Pertinence par rapport aux outils

- Prendre en compte les possibilités des outils disponibles
- Choisir des outils en fonction de textes à traiter :
  - robustesse
  - langue
  - format
- Outils monolingues ou bien multilingues

Exemple de corpus spécialisés

Corpus Menelas :

- genres :
  - manuels
  - comptes rendus d'examens et de traitements
  - lettres aux collègues
- domaine :
  maladies coronariennes

Exemple de corpus spécialisés

Corpus Clef médical :

- genres :
  - compte-rendus d'hospitalisation
  - RMO
  - portail médical (CISMeF)
- domaines :
  - stomatologie
  - néphrologie
  - neurologie
  - etc.

Autres types de corpus

corpus comparables constituent des sélections de textes similaires (langues ou variétés d'une langue)

corpus parallèles sont constitués de documents traduits dans une ou plusieurs langues

corpus alignés les passages correspondants sont réliés (Hansard, EMEA)

corpus de suivi corpus en évolution, corpus glissant

corpus segmentés segmentation en mots, phrases, etc.

corpus étiquetés étiquetage morpho-syntaxique

corpus arborés analyse syntaxique

etc.

Problèmes juridiques

- Information confidentielle, nominale
  la loi Jardé
  - Menelas (CRH : maladies coronariennes)
    secret médical
  - Safir (documents divers : cogénération)
    données confidentielles d'une entreprise (EDF)
- Propriété intellectuelle (forums, tweets...)
  - droits d'auteur
  - droits d'annotateurs de corpus

Problèmes juridiques

Solutions :

- secret médical (Informatique et Liberté)
  anonymisation ou dé-identification :
  - nom du patient
  - service
  - date et lieu de naissance
  - coordonnées du patient
  - coordonnées du service
  - date de consultation, d'hospitalisation
  - nom du médecin

Problèmes juridiques

Solutions :

- propriété intellectuelle :
  - extraits
  - convention
    cession de droits, licence d'utilisation, etc.
- confidentialité vis-à-vis d'une entreprise :
  - convention, achat d'une licence
  - documents " non-sortables "
  - ???

Echantillonnage

- problèmes juridiques
- "équilibrer" en taille les textes
  - "représenter" une diversité maximale de situations de communication
  - ne pas sur-représenter des "lieux" de textes aux caractéristiques particulières
- problèmes :
  comportements hétérogènes des occurrences dans les documents

Documentation des corpus

Types d'informations :

- contexte de production du texte
  auteur, date, taille, format, public visé, thème, objectif, etc.
- contexte de collecte de corpus
  date, responsables, taille, etc.

Enregistrement :

- dans des tables ou tables relationnelles
- encodage XML, BD

Plan

1. Introduction
2. Exemples de terminologies (domaine médical)
3. Constitution de terminologie à partir de corpus
    1. *Extraction de termes*
    2. Extraction de relations

Plan

- Approches pour l'extraction de termes
- Outils pour l'extraction

Introduction

Textes de spécialité : Accès aux informations du domaine (médecine, aviation, électricité, etc.)

Exemple d'application : Extraction d'information à partir de de textes de spécialité (articles scientifiques biomédicaux, dossiers patients, textes de loi, etc.) [Cohen & Demner-Fushman, 2013, Meystre *et al.*, 2008]

Points d'appui :

- Utilisation d'exemples annotés
- Augmentation de la couverture des textes grâce à des ressources terminologiques
    - Thesaurus, nomenclature, glossaire, classification
    - Exemples : MeSH, MedDRA, EPA, IUPAC, Engineering Information thesaurus

Introduction

Exemple

22 yo male, h/o primitive neuroectodermal tumor with mets to **brain**$_{C0006104}$ *and* **spine**$_{C0037949}$, *transferred from* Hospital1, *initially in* Dept1 *and then transferred to the floor.*
He was initially diagnosed with a **thoracic**$_{C0817096}$ *ganglioglioma /resected in 2012. He had* **back**$_{C0004600}$ *pain in 2/04, seen at* Dept2, *and was found to have mets to* **brain**$_{C0006104}$ *and* **spine**$_{C0037949}$.

en gras: termes issus de l'UMLS/ANAT

Introduction

Exemple

22 yo male, h/o primitive **neuroectodermal** *tumor with mets to* **brain**$_{C0006104}$ *and* **spine**$_{C0037949}$*, transferred from* Hospital1*, initially in* Dept1 *and then transferred to the floor.*
He was initially diagnosed with a **thoracic**$_{C0817096}$ *ganglioglioma /resected in 2012. He had* **back**$_{C0004600}$ *pain in 2/04, seen at* Dept2*, and was found to have mets to* **brain**$_{C0006104}$ *and* **spine**$_{C0037949}$*.*

en gras: termes issus de l'UMLS/ANAT

Mais ces ressources sont insuffisantes

[Bodenreider *et al.*, 2002, McCray *et al.*, 2002]

Introduction

Exemple

22 yo male, h/o primitive **neuroectodermal** *tumor with mets to* **brain**$_{C0006104}$ *and* **spine**$_{C0037949}$*, transferred from* Hospital1*, initially in* Dept1 *and then transferred to the floor.*
He was initially diagnosed with a **thoracic**$_{C0817096}$ *ganglioglioma /resected in 2012. He had* **back**$_{C0004600}$ *pain in 2/04, seen at* Dept2*, and was found to have mets to* **brain**$_{C0006104}$ *and* **spine**$_{C0037949}$*.*

en gras: termes issus de l'UMLS/ANAT

Mais ces ressources sont insuffisantes

[Bodenreider *et al.*, 2002, McCray *et al.*, 2002]

Il est souvent nécessaire :

- d'adapter les ressources terminologiques aux textes à traiter
  (problème de couverture, d'adéquation, etc.)

Introduction

Exemple

22 yo male, h/o primitive **neuroectodermal** *tumor with mets to* **brain**$_{C0006104}$ *and* **spine**$_{C0037949}$*, transferred from* Hospital1*, initially in* Dept1 *and then transferred to the floor.*
He was initially diagnosed with a **thoracic**$_{C0817096}$ *ganglioglíom /resected in 2012. He had* **back**$_{C0004600}$ *pain in 2/04, seen at* Dept2*, and was found to have mets to* **brain**$_{C0006104}$ *and* **spine**$_{C0037949}$*.*

en gras: termes issus de l'UMLS/ANAT

Mais ces ressources sont insuffisantes

[Bodenreider *et al.*, 2002, McCray *et al.*, 2002]

Il est souvent nécessaire :

- d'adapter les ressources terminologiques aux textes à traiter
  (problème de couverture, d'adéquation, etc.)
- de créer des ressources spécifiques
  (pas de ressources adaptées qui décrivent les informations visées, etc.)

Disposer de ressources terminologiques adaptées à la tâche

- Identifier des variantes des termes

[Jacquemin, 1997, Nenadic *et al.*, 2004, Spasić *et al.*, 2013]

mets to brain and spine $\rightarrow$ **spine metastases**$_{C0684550}$

Inapplicable lorsqu'on ne dispose pas de terminologie ou que les types sémantiques des entités recherchées ne sont pas présents dans les terminologies disponibles

Disposer de ressources terminologiques adaptées à la tâche

- Identifier des variantes des termes

  [Jacquemin, 1997, Nenadic *et al.*, 2004, Spasić *et al.*, 2013]

  mets to brain and spine → **spine metastases**$_{C0684550}$

  Inapplicable lorsqu'on ne dispose pas de terminologie ou que les types sémantiques des entités recherchées ne sont pas présents dans les terminologies disponibles

- Extraire les termes potentiels

  [Cabré *et al.*, 2001, Pazienza *et al.*, 2005]

  et regrouper ces termes grâce à des méthodes d'acquisition de relations sémantiques [Grabar & Hamon, 2004]

Exemple

22 yo [*male*], *h/o* [*primitive neuroectodermal tumor*] *with* [*mets*] *to* [**brain**$_{C0006104}$] *and* [**spine**$_{C0037949}$], *transferred from* Hospital1, *initially in* Dept1 *and then transferred to the* [*floor*]. *He was initially diagnosed with a* [**thoracic**$_{C0817096}$ *ganglioglioma*] */resected in 2012. He had* [**back**$_{C0004600}$ **pain**$_{C0004604}$] *in 2/04, seen at* Dept2, *and was found to have* [*mets*] *to* [**brain**$_{C0006104}$] *and* [**spine**$_{C0037949}$].

en gras: termes issus de l'UMLS/ANAT – [box] : termes candidats

Un terme ou non ?

- Qu'est-ce qui n'est pas un terme ?
- Qu'est-ce qu'un terme ?

Un terme ou non ?

- *Dermatose acantholytique*
- *Crampes de l'abdomen*
- *Prédisposition accident*
- *Dettes à recouvrer*
- *Acupuncture, traitement*
- *Acides acétiques*
- *Acétiques, acides*
- *Syndrome Adams Strokes*
- *Adams Strokes, syndrome*

Un terme ou non ?

- *Onzième paire crânienne*
- *Huitième paire crânienne, maladie*
- *Troubles de l'adaptation avec perturbation mixte des émotions et des conduites*
- *Malformations induites par les composés chimiques*
- *Syndrome de sécrétion inappropriée d'hormone de croissance*
- *Nucléoside-2',3'-cyclic-phosphate 3'-nucléotido-hydrolase*
- *Désoxyribonucléase (ATP-and D-adénosyl-L-méthionine-dépendante)*

Un terme ou non ?

A. Rey. La terminologie. Noms et notions. Que sais-je ?

La terminologie exclue :

- marques d'énonciation :
  - pronoms personnels
  - adjectifs possessifs
  - adverbes de temps et de lieu
- mots " grammaticaux "
- verbes (sauf si assimilables à un nom)

Où sont les termes ?

En cas d'intolérance aux inhibiteurs de l'enzyme de conversion, dans le cadre de l'insuffisance cardiaque chronique congestive, l'essai des Veterans (V-HeFT II) a montré la possibilité d'utiliser comme traitement substitutif l'association hydralazine (37,5 mg/j) - dinitrate d'isosorbide (20 mg/j). Les antagonistes des récepteurs de l'angiotensine II (losartan) mis sur le marché avec l'indication hypertension artérielle sont actuellement en cours d'étude pour évaluer leur effet thérapeutique en termes de morbidité ou mortalité dans l'insuffisance cardiaque.

Où sont les termes ?

*En cas d'***intolérance aux inhibiteurs de l'enzyme de conversion**, *dans le cadre de l'***insuffisance cardiaque chronique congestive**, *l'***essai des Veterans** *(***V-HeFT II***) a montré la possibilité d'utiliser comme* **traitement substitutif** *l'association* **hydralazine** *(37,5 mg/j) -* **dinitrate d'isosorbide** *(20 mg/j). Les antagonistes des récepteurs de l'angiotensine II (***losartan***) mis sur le marché avec l'indication* **hypertension artérielle** *sont actuellement en cours d'étude pour évaluer leur* **effet thérapeutique** *en termes de* **morbidité** *ou* **mortalité** *dans l'***insuffisance cardiaque***.*

Où sont les termes ?

Combined action of two transcription factors regulates genes encoding spore coat proteins of Bacillus subtilis.
During sporulation of Bacillus subtilis, spore coat proteins encoded by cot genes are expressed in the mother cell and deposited on the forespore. Transcription of the cotB, cotC, and cotX genes by final sigma(K) RNA polymerase is activated by a small, DNA-binding protein called GerE. The promoter region of each of these genes has two GerE binding sites. 5' deletions that eliminated the more upstream GerE site decreased expression of lacZ fused to cotB and cotX by ...

Où sont les termes ?

Combined action of two **transcription factors** regulates genes encoding **spore coat proteins** of Bacillus subtilis.
During **sporulation of Bacillus subtilis**, **spore coat proteins** encoded by **cot genes** are expressed in the **mother cell** and deposited on the **forespore**. **Transcription of the cotB**, cotC, and **cotX genes** by final **sigma(K) RNA polymerase** is activated by a small, **DNA-binding protein** called **GerE**. The **promoter region** of each of these **genes** has two **GerE binding sites**. 5' deletions that eliminated the more **upstream GerE site** decreased **expression of lacZ** fused to cotB and cotX by ...

Vers une acquisition automatique

Terminologie descriptive

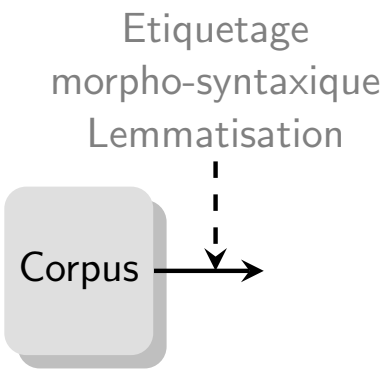
Traditionnellement : Méthodes semi-automatiques d'acquisition terminologique destinées à aider les terminologues à construire des terminologies

1. Constitution d'une liste de termes candidats
2. Mise en relation des termes candidats
3. Validation par un terminologue des informations extraites
   $\rightarrow$ Définition de fiches terminologiques

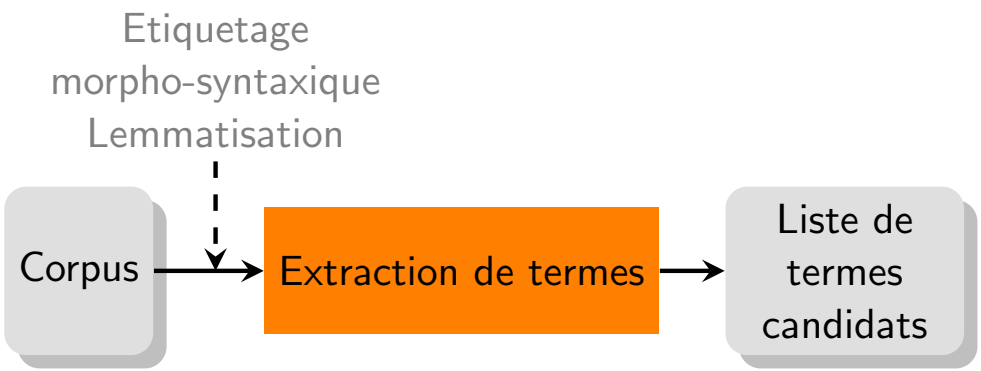
Processus de construction d'une terminologie

Corpus

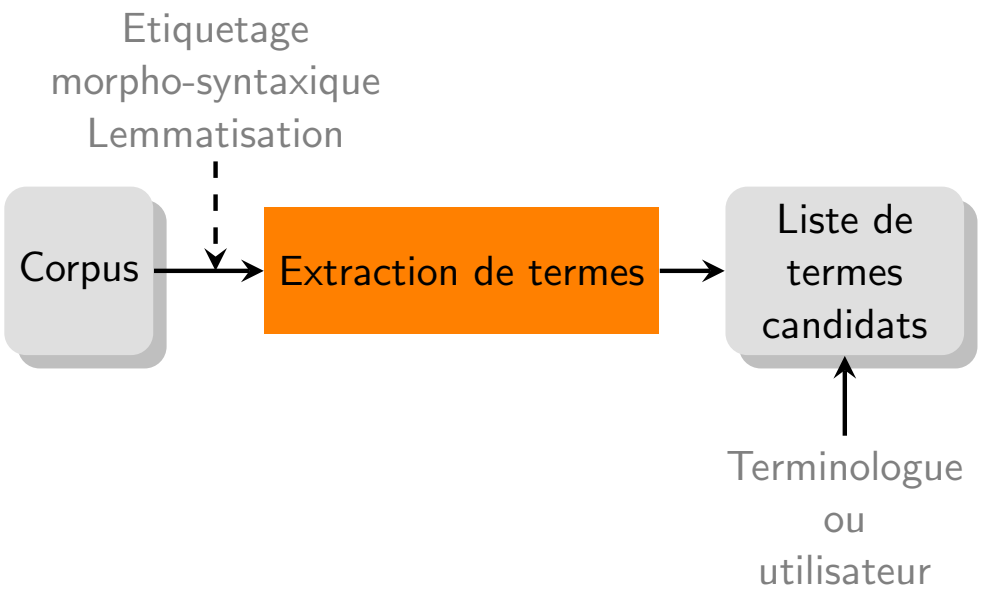
Processus de construction d'une terminologie



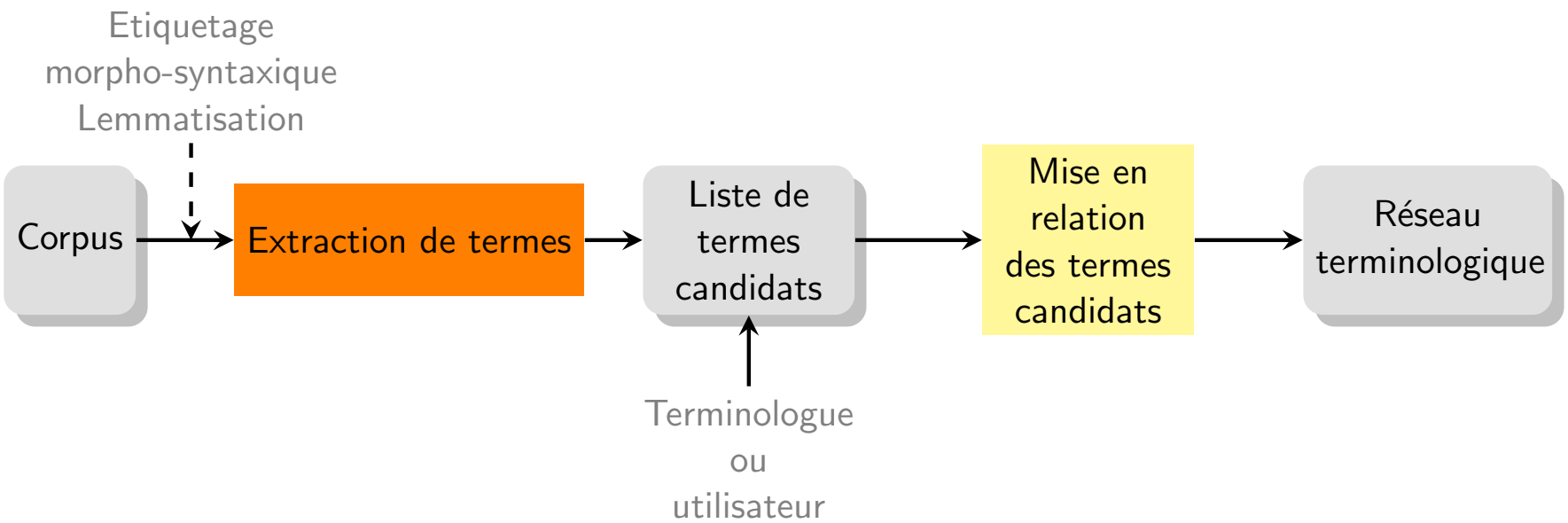
Processus de construction d'une terminologie



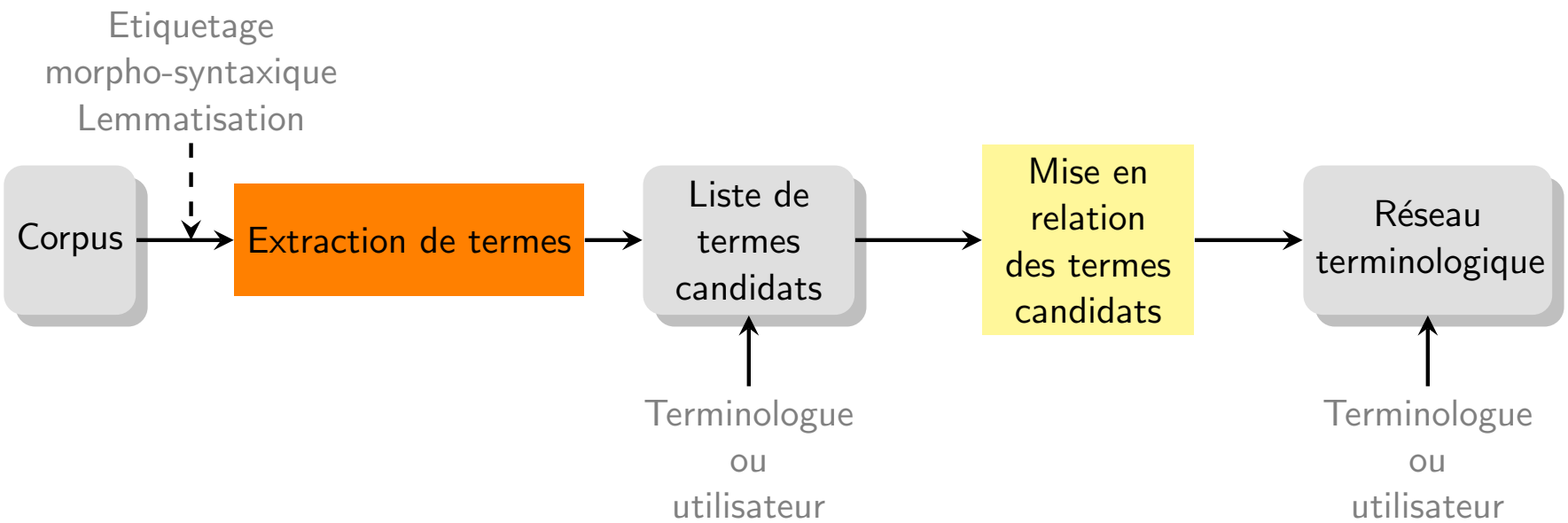
Processus de construction d'une terminologie



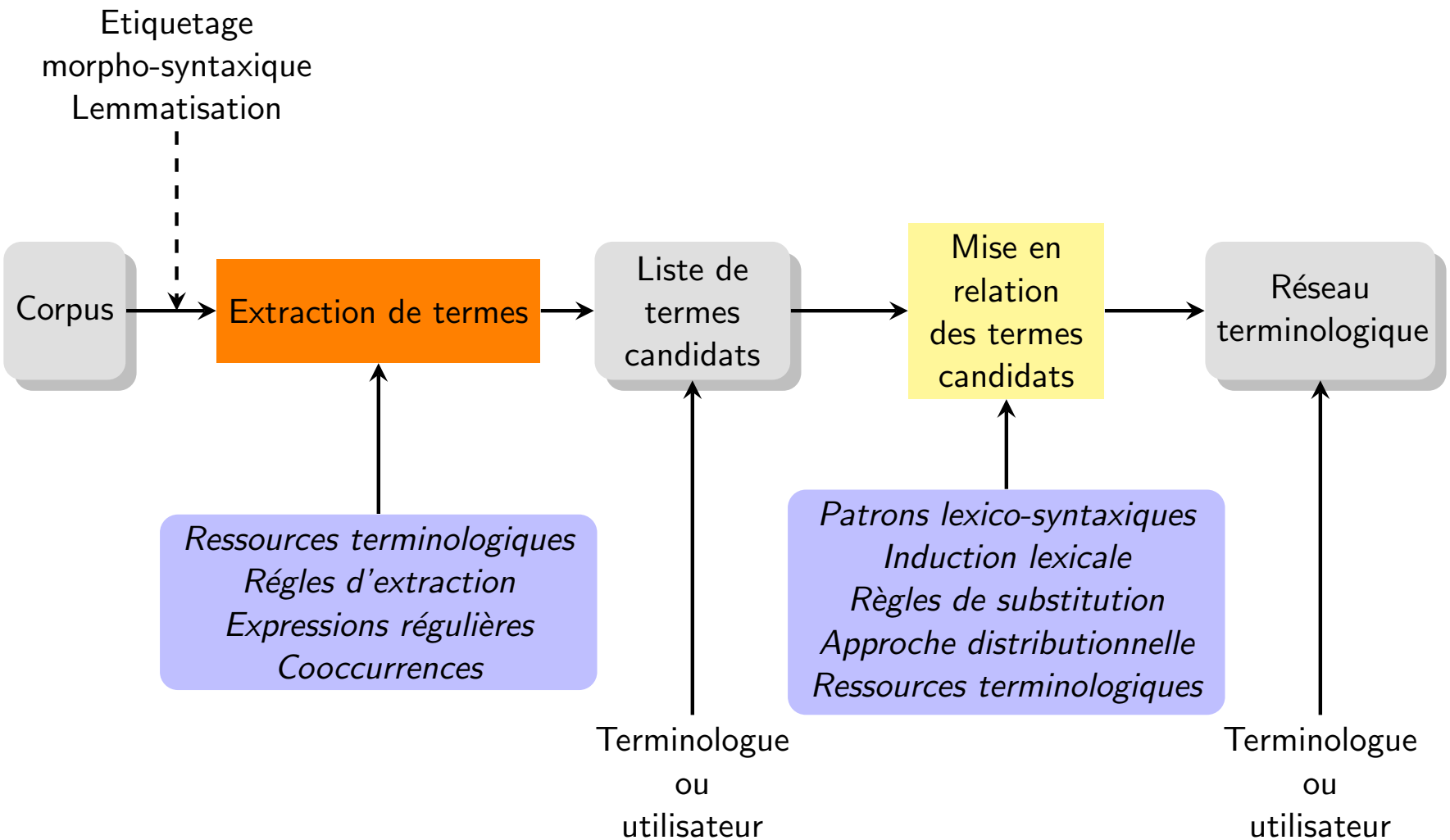
Processus de construction d'une terminologie



Processus de construction d'une terminologie



Processus de construction d'une terminologie



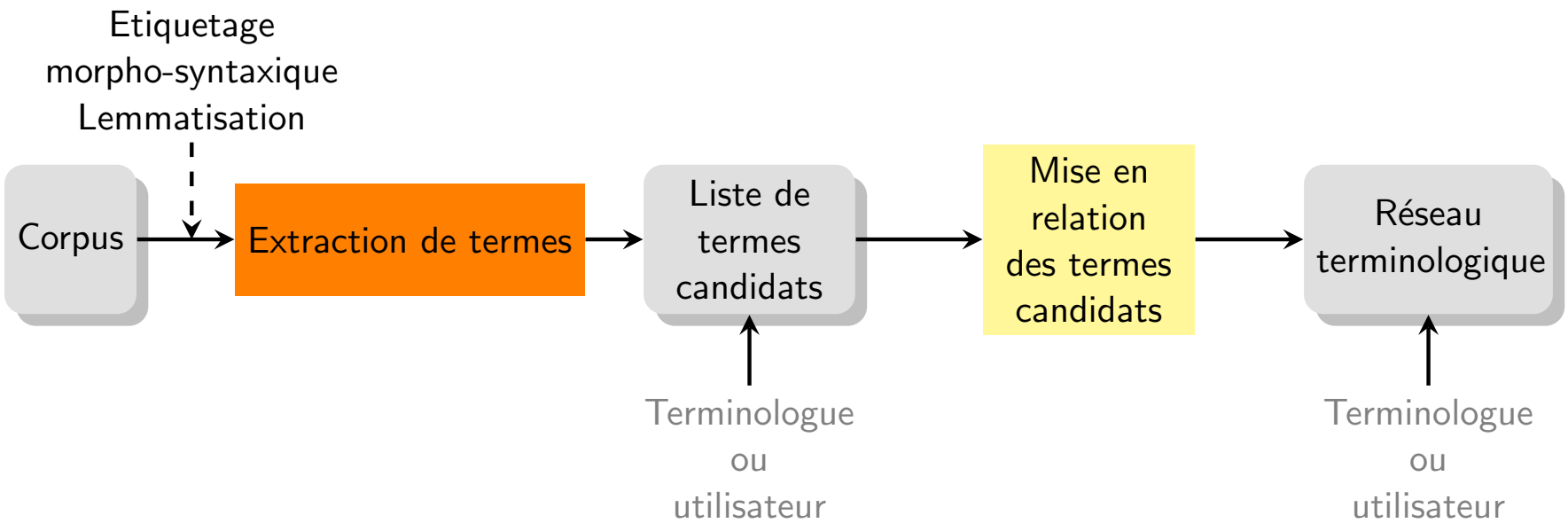
Vers une acquisition automatique

Principalement des approches linguistiques à base de règles (prise en compte des contraintes théoriques de constitution de terminologies)

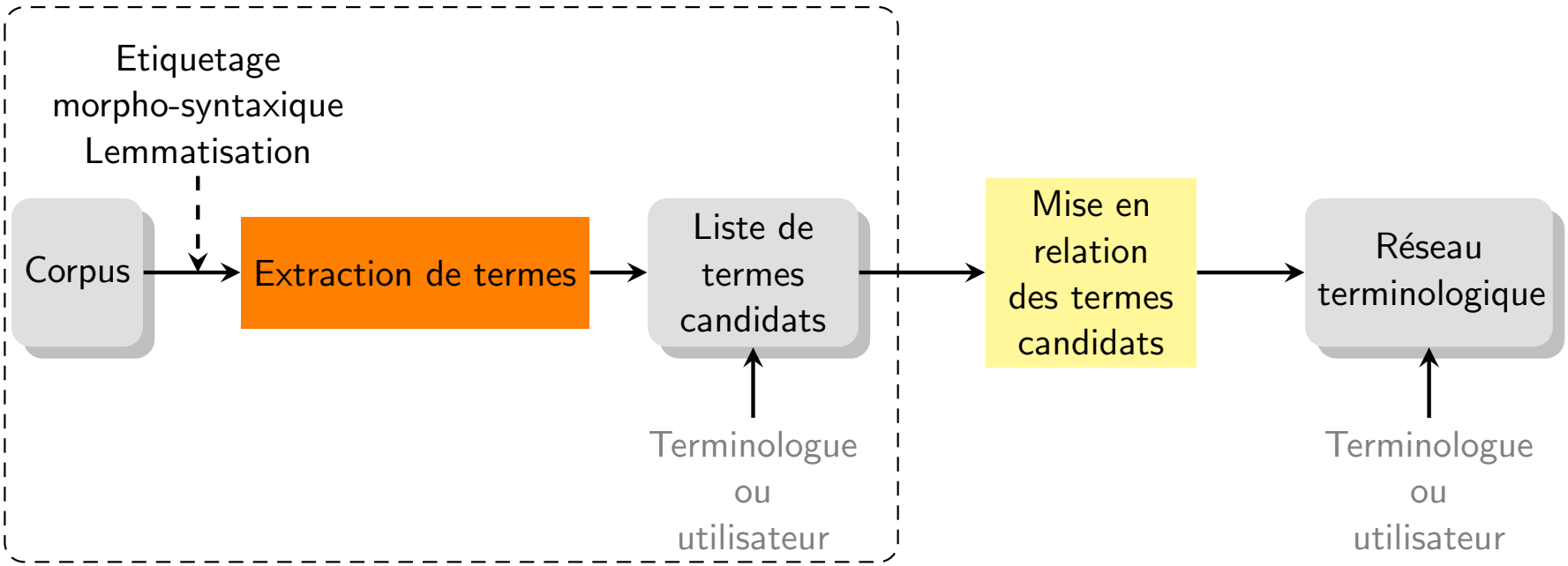
Utilisation de ces méthodes

- de manière complètement automatique
- pour l'adaptation de ces ressources
- dans le cadre d'applications réelles

Extraction terminologique



Extraction terminologique



Reconnaissance vs. extraction de termes

- Reconnaissance : Projection des termes issus d'une terminologie sur un texte
  Utilisation de méthodes plus ou moins complexes (projection directe, calcul de variantes terminologiques, distance sémantique, etc.)

- Extraction : Découverte des termes directement dans le corpus
  Identification des syntagmes (nominaux) pouvant être des termes
  Calcul de :
  - la cohésion de leurs composants (*unithood*)
  - leur caractère terminologique (*termhood*)

[Kageura & Umino, 1996]

Processus de construction d'une terminologie

Textes

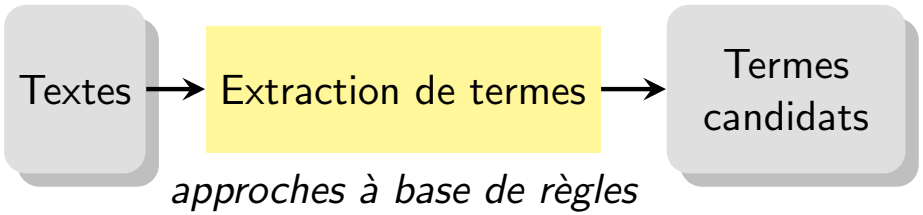
22 yo male , h / o primitive neuroectodermal tumor with mets to brain and spine , transferred from Hospital1 , initially in Dept1 and then transferred to the floor .
He was initially diagnosed with a thoracic gangliogliom / resected in 2012 .
He had back pain in 2 / 04 , seen at Dept2 , and was found to have mets to brain and spine .

Processus de construction d'une terminologie

lemmatisation
+ POS tagging

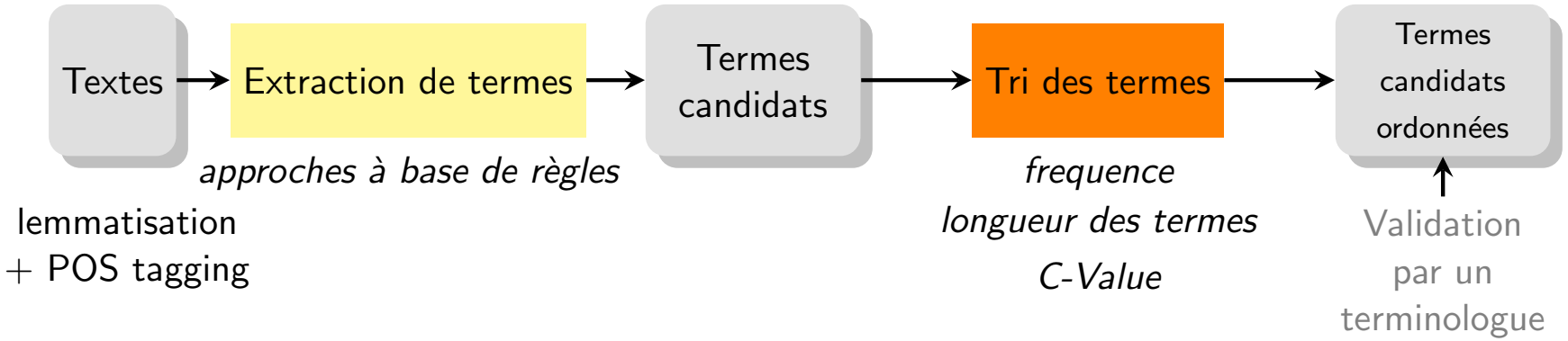
22$_{CD}$ yo$_{JJ}$ male$_{NN}$,$_{,}$ h$_{NN}$ /$_{SYM}$ o$_{NN}$ primitive$_{JJ}$ neuroectodermal$_{JJ}$ tumor$_{NN}$ with$_{IN}$ met$_{NNS}$ to$_{TO}$ brain$_{NN}$ and$_{CC}$ spine$_{NN}$,$_{,}$ transfer$_{VBN}$ from$_{IN}$ Hospital1$_{NNP}$,$_{,}$ initially$_{RB}$ in$_{IN}$ Dep1$_{NNP}$ and$_{CC}$ then$_{RB}$ transfer$_{VBN}$ to$_{TO}$ the$_{DT}$ floor$_{NN}$.$_{.}$
He$_{PRP}$ be$_{VBD}$ initially$_{RB}$ diagnose$_{VBN}$ with$_{IN}$ a$_{DT}$ thoracic$_{JJ}$ ganglioglom$_{NN}$ /$_{SYM}$ resecte$_{VBN}$ in$_{IN}$ 2012$_{CD}$.$_{.}$
He$_{PRP}$ have$_{VBD}$ back$_{JJ}$ pain$_{NN}$ in$_{IN}$ 2$_{CD}$ /$_{SYM}$ 04$_{CD}$,$_{,}$ see$_{VBN}$ at$_{IN}$ Dept2$_{NNP}$,$_{,}$ and$_{CC}$ be$_{VBD}$ find$_{VBN}$ to$_{TO}$ have$_{VB}$ met$_{NNS}$ to$_{TO}$ brain$_{NN}$ and$_{CC}$ spine$_{NN}$.$_{.}$

Processus de construction d'une terminologie



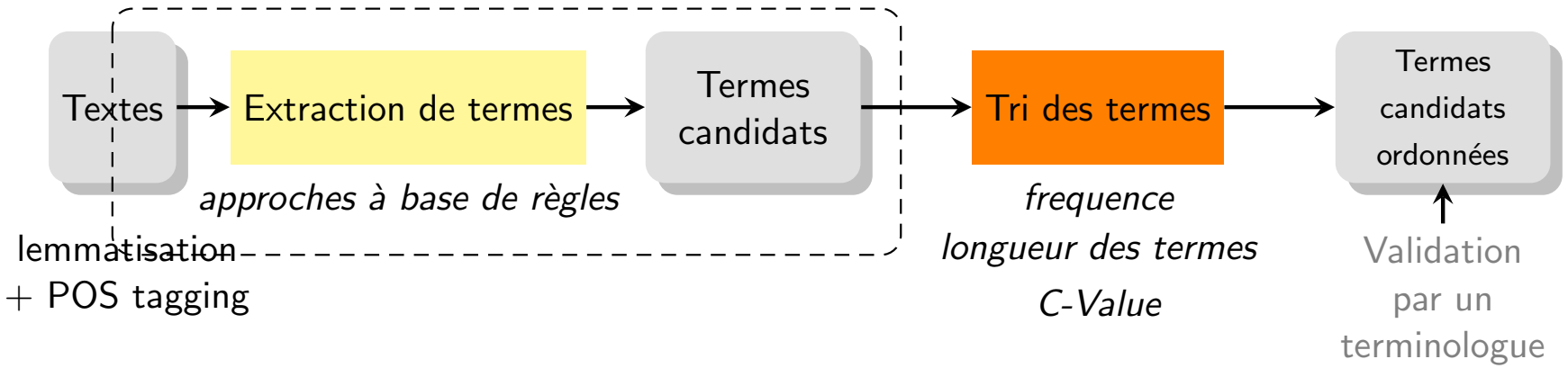
yo male	*thoracic gangliogliom*
h	*back pain*
o	*mets*
primitive neuroectodermal tumor	*brain*
mets	*spine*
brain	*floor*
spine	
...	

Processus de construction d'une terminologie



	f	l	Cv_1		f	l	Cv_1
yo male	*1*	*1*	*1.58*	*spine*	*2*	*1*	*2*
h	*1*	*1*	*1*	*floor*	*1*	*1*	*1*
o	*1*	*1*	*0*	*thoracic gangliogliom*	*1*	*2*	*1.58*
mets	*2*	*1*	*2*	*back pain*	*1*	*2*	*1.58*
brain	*2*	*1*	*2*				
primitive neuroectodermal tumor					*1*	*3*	*2.32*
...							

Processus de construction d'une terminologie



	f	l	Cv_1		f	l	Cv_1
yo male	*1*	*1*	*1.58*	*spine*	*2*	*1*	*2*
h	*1*	*1*	*1*	*floor*	*1*	*1*	*1*
o	*1*	*1*	*0*	*thoracic gangliogliom*	*1*	*2*	*1.58*
mets	*2*	*1*	*2*	*back pain*	*1*	*2*	*1.58*
brain	*2*	*1*	*2*				
primitive neuroectodermal tumor					*1*	*3*	*2.32*
...							

Approches pour l'extraction de termes

Amorcer

- découpage de la phrase sur
  les frontières syntaxiques des syntagmes terminologiques
  - pronoms, verbes conjugués
  - prépositions
  - coordination
  - ponctuation
- repérage de connecteurs grammaticaux
  de, de l', du, etc.
- repérage d'ancres lexicales
  mots " centraux " déjà connus

Approches pour l'extraction de termes

Extraire

- recherche de segments répétés dans une fenêtre de n mots
- recherche de patrons syntaxiques de groupes nominaux
- recherche de patrons syntaxiques de groupes nominaux et adjectivaux
- application de patrons syntaxiques de bitermes
- repérage de syntagmes répétés autour de connecteurs grammaticaux
- repérage de syntagmes répétés autour d'ancres lexicales

Approches pour l'extraction de termes

Affiner

- décomposition en syntagmes minimaux
- filtres statistiques
- filtres lexicaux
- application de règles de variation
- fusion de variantes

Approches pour l'extraction de termes

- Expressions régulières et filtrage statistique
  ATR & C-Value [Frantzi *et al.*, 2000]
- Termes et variantes
  - Bitermes et variantes & mesures statistiques (ACABIT) [Daille, 1995]
  - Grammaire de termes et méta-règles pour l'appariement des variantes terminologiques (Faster) [Jacquemin, 1997]
- Approche contrastive
  Extraction des termes autour de pivôts lexicaux spécialisés (TermoStat) [Drouin, 2002]
- Analyse syntaxique et désambiguïsation endogène

  Analyse superficielle à base de règles, en cascade (Lexter, Syntex) [Bourigault *et al.*, 2005]

  Analyse superficielle à base de patrons minimaux ($\mathrm{Y_A T_E A}$) [Aubin & Hamon, 2006]

Acabit

Béatrice Daille (1995), Repérage et extraction de terminologie par une approche mixte statistique et linguistique. *TAL* 36(1-2), p. 101-118.

- Approche mixte linguistique et statistique
- Bitermes et leurs variantes
- Extraction de candidats termes à partir d'un corpus préalablement étiqueté et désambiguïsé

Acabit

L'acquisition terminologique dans Acabit se déroule en deux étapes :

1. Analyse linguistique et regroupement de variantes :
   - Corpus étiqueté
   - Transducteurs pour la recherche de séquences nominales
   - Extraction de candidats termes :
     - N ADJ : *station terrienne*
     - N_1 PREP N_2 : *liaison par satellite*
     - N_1 N_2 : *diode tunnel*
   - Décomposition en candidats termes binaires :

     réseau de transit à satellite
     → *réseau de transit*
     → *réseau à satellite*

Acabit

L'acquisition terminologique dans Acabit se déroule en deux étapes :

2. Filtrage statistique :
   - Mesures statistiques pour le tri de candidats termes binaires
   - Calcul de scores et de distances sur les composants des candidats termes basés sur les fréquences
   - log-likelihood ratio (Dunning, 1993)
     le mieux pour retenir les termes candidats sans être sensible aux fréquences

Lexter

Didier Bourigault (1993), Analyse syntaxique locale pour le repérage de termes complexes dans un texte. *TAL*, p. 105-117
Analyse endogène (pas de connaissance du domaine)

- Extraction de candidats termes à partir d'un corpus étiqueté et désambiguïsé
- Analyse syntaxique de surface
- Repérage et analyse des syntagmes nominaux
- Organisation de l'ensemble des candidats termes en un réseau

Lexter

L'acquisition des termes est effectuée en trois étapes :

1. Extraction de syntagmes nominaux maximaux
2. Décomposition de syntagmes maximaux
3. Module de structuration

Lexter

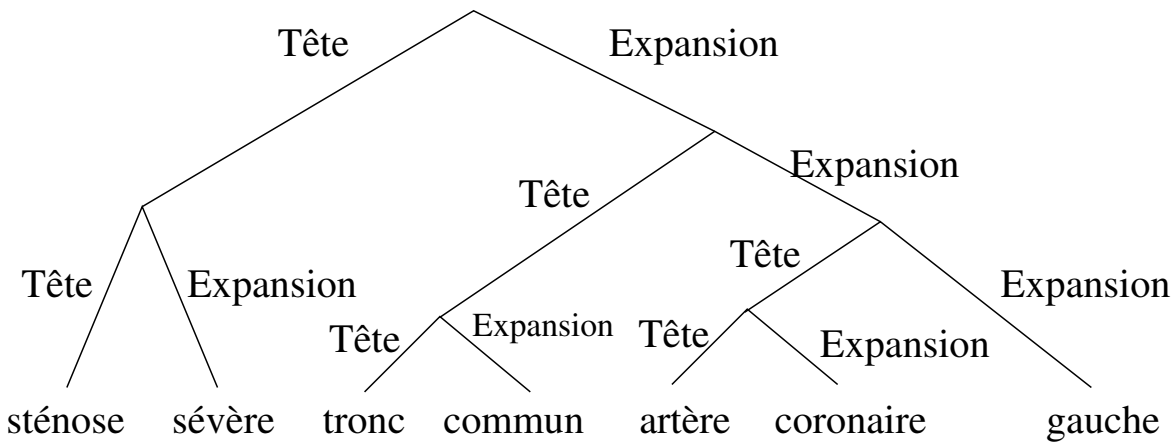
1. Extraction de syntagmes nominaux maximaux

- Repérage de frontières syntaxiques
  verbes conjugués, pronoms, conjonctions de subordination, ...
- Extraction de syntagmes nominaux maximaux
- Apprentissage endogène sur corpus
- Informations de sous-catégorisation des noms et des adjectifs, propres aux corpus
- Résolution de cas d'ambiguïté de rattachement prépositionnel
- Dans un corpus, le nom *pression* sous-catégorise la préposition *à* :
  - *pression à l'aspiration*
  - *pression au refoulement*

Lexter

2. Décomposition de syntagmes maximaux

- Décomposition récursive de syntagmes nominaux maximaux
- Tête et expansion syntaxiques
- *sténose sévère du tronc commun de l'artère coronaire gauche*



Lexter

2. Décomposition de syntagmes maximaux

- Apprentissage endogène sur corpus
- Ambiguïté de rattachement au sein de ces groupes nominaux
- Candidats termes :
  - syntagmes maximaux
  - leurs constituants

Lexter

3. Module de structuration

- Construction d'un réseau de candidats termes
- Relation de chaque candidat à ceux dont il est tête ou expansion
  sténose / *sténose sévère*
  ...
- Calcul d'un coefficient de productivité
  densité du réseau autour d'un candidat terme

Lexter

Exemple de sortie Lexter

- En entrée :

  <Prep>En <NomFS>présence <Prep>de <NomFS>sténose <Adj?S>sévère <Prep>de <DetMS>le <NomMS>tronc <Adj?S>commun <Prep>de <Det?S>l' <NomFS>artère <Adj?S>coronaire <Adj?S>gauche <Typo>, <Det?S>on <Pro>se <VCONJ>contente <Prep>d' <Det>un <Nom?S>minimum <Prep>d' <NomFP>injections <Typo>,

- Extraction de candidats termes :
  $\longrightarrow$ *(sténose sévère) du (tronc commun de l'((artère coronaire) gauche))*
  $\longrightarrow$ *minimum d'injections*

TermoStat

(Drouin 2002)

- Recours à des tests statistiques :
  - Comparaison du lexique du corpus (de spécialité) avec un corpus de référence (général)
  - Calcul d'un indice de spécificité (Lebart et Salem 1994) associé à chaque mot
- Identification de pivots lexicaux spécialisés (PLS)
  - Identifier les termes simples les plus représentatifs du corpus de spécialité par contraste avec un corpus général

TermoStat

Exemple (identification de PLS)

For/IN Dual/JJ MSA/NNP sites/NNS (/(line/NN sites/NNS with/IN high/JJ OADM/NNP counts/NNS)/SYM shown/VBN in/IN Figure/NN 4/CD -/: 12/CD ,/, the/DT signal/NN flow/NN is/VBZ the/DT same/JJ except/IN that/DT a/DT second/JJ MSA/NNP (/(DSCM/NNP or/CC OADM/NNP filter/NN)/SYM is/VBZ placed/VBN between/IN the/DT Booster18/NNP and/CC Booster21/NNP circuit/NN packs/NNS ./.

(exemple issu de Drouin 2002)

TermoStat

Exemple (identification de PLS)

For/IN Dual/JJ **MSA/NNP sites/NNS** *(/(* **line/NN sites/NNS** *with/IN high/JJ* **OADM/NNP counts/NNS** *)/SYM shown/VBN in/IN Figure/NN 4/CD -/: 12/CD ,/, the/DT signal/NN* **flow/NN** *is/VBZ the/DT same/JJ except/IN that/DT a/DT second/JJ* **MSA/NNP** *(/(* **DSCM/NNP** *or/CC* **OADM/NNP filter/NN** *)/SYM is/VBZ placed/VBN between/IN the/DT* **Booster18/NNP** *and/CC* **Booster21/NNP circuit/NN packs/NNS** *./.*

(exemple issu de Drouin 2002)

TermoStat

- PLS : amorce pour l'extraction de termes
  Utilisation des frontières de termes (Bourigault 1994) pour extraire les termes candidats :
  - Frontière à droite : le PLS (tête du terme)
  - Frontière à gauche : un élément du texte ne pouvant apparaître dans un terme

  Elimination des termes candidats construits à partir de têtes moins pertinentes pour le domaine

TermoStat

Exemple (extraction des termes)

For/IN Dual/JJ MSA/NNP sites/NNS (/(line/NN sites/NNS with/IN high/JJ OADM/NNP counts/NNS)/SYM shown/VBN in/IN Figure/NN 4/CD -/: 12/CD ,/, the/DT signal/NN flow/NN is/VBZ the/DT same/JJ except/IN that/DT a/DT second/JJ MSA/NNP (/(DSCM/NNP or/CC OADM/NNP filter/NN)/SYM is/VBZ placed/VBN between/IN the/DT Booster18/NNP and/CC Booster21/NNP circuit/NN packs/NNS ./.

(exemple issu de Drouin 2002)

TermoStat

Exemple (extraction des termes)

For/IN Dual/JJ **MSA/NNP** **sites/NNS** *((* **line/NN** **sites/NNS** *with/IN high/JJ* **OADM/NNP** **counts/NNS** *)/SYM shown/VBN in/IN Figure/NN 4/CD -/: 12/CD ,/, the/DT signal/NN* **flow/NN** *is/VBZ the/DT same/JJ except/IN that/DT a/DT second/JJ* **MSA/NNP** *((* **DSCM/NNP** *or/CC* **OADM/NNP** **filter/NN** *)/SYM is/VBZ placed/VBN between/IN the/DT* **Booster18/NNP** *and/CC* **Booster21/NNP** **circuit/NN** **packs/NNS** *./.*

(exemple issu de Drouin 2002)

TermoStat

Exemple (extraction des termes)

Dual/JJ **MSA/NNP** **sites/NNS** **line/NN**
sites/NNS *high/JJ* **OADM/NNP** **counts/NNS**

flow/NN *same/JJ*
second/JJ **MSA/NNP** **DSCM/NNP**
OADM/NNP **filter/NN**
Booster18/NNP **Booster21/NNP**
circuit/NN **packs/NNS**

(exemple issu de Drouin 2002)

TermoStat

Exemple (extraction des termes)

[Dual/JJ **MSA/NNP** *]* **sites/NNS** *[* **line/NN** *]*
sites/NNS *[high/JJ* **OADM/NNP** *]* **counts/NNS**

[**flow/NN** *]* *same/JJ*
second/JJ [**MSA/NNP** *]* *[* **DSCM/NNP** *]*
[**OADM/NNP** *]* **filter/NN**
[**Booster18/NNP** *]*
*[***Booster21/NNP** *]* **circuit/NN** **packs/NNS**

(exemple issu de Drouin 2002)

TermoStat

Exemple (extraction des termes)

[[Dual/JJ **MSA/NNP** *]* **sites/NNS** *]* *[[* **line/NN** *]* **sites/NNS** *]* *[[high/JJ* **OADM/NNP** *]* **counts/NNS** *]*

[**flow/NN** *]* *same/JJ* *second/JJ [* **MSA/NNP** *]* *[* **DSCM/NNP** *]* *[[* **OADM/NNP** *]* **filter/NN** *]* *[* **Booster18/NNP** *]* *[* *[***Booster21/NNP** *]* **circuit/NN** *]* **packs/NNS**

(exemple issu de Drouin 2002)

TermoStat

Exemple (extraction des termes)

[[Dual/JJ **MSA/NNP** *]* **sites/NNS** *] [[* **line/NN** *]* **sites/NNS** *] [[high/JJ* **OADM/NNP** *]* **counts/NNS** *]*

[**flow/NN** *] same/JJ second/JJ [* **MSA/NNP** *] [* **DSCM/NNP** *]*
[[**OADM/NNP** *]* **filter/NN** *]*
[**Booster18/NNP** *] [[*
*[***Booster21/NNP** *]* **circuit/NN** *]* **packs/NNS** *]*

(exemple issu de Drouin 2002)

TermoStat

- Tri des termes candidats suivant un *indice terminologique* (*iTer*)
  Prise en compte de
  - la fréquence
  - la longueur du terme
  - la fréquence de la tête potentielle d'un terme candidat

ATR

(Frantzi *et al* 2000)

Combinaison d'informations linguistiques et statistiques

- Filtrage linguistique : séquences de mots caractéristiques des termes, composées à partir de catégories morpho-syntaxiques

  `Noun+Noun`
  `(Adj|Noun)+Noun`
  `((Adj|Noun)+|((Adj|Noun)*(NounPrep)?)(Adj|Noun)*)Noun`

- Anti-dictionnaire : *great*, *numerous*, *several*, *year*, *just*, *good*, etc.
- Filtrage statistique : *C-value*
  - Prend en compte des informations statistiques associées aux termes
  - Mesure l'indépendance des termes
  - Privilégie les termes longs et qui ne sont pas des composant d'autres termes

ATR

(Frantzi *et al* 2000)

$$C-value(t) = \begin{cases} log_2(|t|) \times f(t) \text{ si } t \text{ n'est pas inclus dans un terme} \\ log_2(|t|) \times (f(t) - \frac{1}{P(T_t)} \sum_{t' \in T_t} f(t')) \text{ sinon} \end{cases}$$

- fréquence du terme ($f(t)$)
- nombre de mots du terme ($|t|$)
- fréquence du terme comme composant d'un terme plus grand ($f(t')$)
- T_t ensemble des termes incluant t
- nombre de termes plus grands incluant le terme ($P(T_t)$)

Variante : *NC-value* (Maynard et Ananiadou 2001) – prise en compte des termes d'un thesaurus par calcul d'une distance sémantique

YaTeA

Yet Another Term ExtrActor
(Aubin et Hamon, 2006)

- Extration de termes sur des textes français et anglais
- Analyse syntaxique superficielle (Tête / Modifeur) à l'aide
  - de patrons minimaux appliqués récursivement
  - de l'apprentissage endogène
- Rejet des groupes nominaux non analysables
- Association de mesures statistiques (Fréquences, C-Value1, C-Value*, etc.) [Hamon *et al.*, 2014]

- Module CPAN `http://search.cpan.org/~thhamon/Lingua-YaTeA/`
- Développement dans le cadre du projet ALVIS
- Description de l'analyse à partir de fichiers de configuration
  Possibilité d'adaptation à un domaine : BioYaTeA [Golik *et al.*, 2013]

YaTeA (2)

- Identification de groupes nominaux à partir de frontières morpho-syntaxiques

  22$_{CD}$ yo$_{JJ}$ male$_{NN}$,$_{,}$ h$_{NN}$/$_{SYM}$o$_{NN}$ primitive$_{JJ}$ neuroectodermal$_{JJ}$ tumor$_{NN}$ with$_{IN}$ mets$_{NNS}$ to$_{TO}$ brain$_{NN}$ and$_{CC}$ spine$_{NN}$,$_{,}$ transferred$_{VBN}$ from$_{IN}$ Hospital1$_{NNP}$,$_{,}$ initially$_{RB}$ in$_{IN}$ Dept1$_{NNP}$ and$_{CC}$ then$_{RB}$ transferred$_{VBN}$ to$_{TO}$ the$_{DT}$ floor$_{NN}$.$_{.}$ He$_{PRP}$ was$_{VBD}$ initially$_{RB}$ diagnosed$_{VBN}$ with$_{IN}$ a$_{DT}$ thoracic$_{JJ}$ ganglioglioma$_{NN}$ /$_{/}$ resected$_{VBN}$ in$_{IN}$ 2012$_{CD}$.$_{.}$ He$_{PRP}$ had$_{VBD}$ back$_{JJ}$ pain$_{NN}$ in$_{in}$ 2$_{CD}$/$_{SYM}$04$_{CD}$,$_{,}$ seen$_{VBN}$ at$_{IN}$ Dept2$_{NNP}$,$_{,}$ and$_{CC}$ was$_{be}$ found$_{VBN}$ to$_{TO}$ have$_{VB}$ mets$_{NNS}$ to$_{TO}$ brain$_{NN}$ and$_{CC}$ spine$_{NN}$.$_{.}$

YaTeA (2)

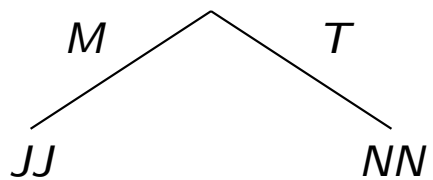
- Identification de groupes nominaux à partir de frontières morpho-syntaxiques

22_{CD} yo_{JJ} $male_{NN}$ $,_{,}$ h_{NN} $/_{SYM}$ o_{NN} $primitive_{JJ}$ $neuroectodermal_{JJ}$ $tumor_{NN}$ $with_{IN}$ $mets_{NNS}$ to_{TO} $brain_{NN}$ and_{CC} $spine_{NN}$ $,_{,}$ $transferred_{VBN}$ $from_{IN}$ $Hospital1_{NNP}$ $,_{,}$ $initially_{RB}$ in_{IN} $Dept1_{NNP}$ and_{CC} $then_{RB}$ $transferred_{VBN}$ to_{TO} the_{DT} $floor_{NN}$ $._{.}$ He_{PRP} was_{VBD} $initially_{RB}$ $diagnosed_{VBN}$ $with_{IN}$ a_{DT} $thoracic_{JJ}$ $ganglioglioma_{NN}$ $/_{/}$ $resected_{VBN}$ in_{IN} 2012_{CD} $._{.}$ He_{PRP} had_{VBD} $back_{JJ}$ $pain_{NN}$ in_{in} 2_{CD} $/_{SYM}$ 04_{CD} $,_{,}$ $seen_{VBN}$ at_{IN} $Dept2_{NNP}$ $,_{,}$ and_{CC} was_{be} $found_{VBN}$ to_{TO} $have_{VB}$ $mets_{NNS}$ to_{TO} $brain_{NN}$ and_{CC} $spine_{NN}$ $._{.}$

YATEA (3)

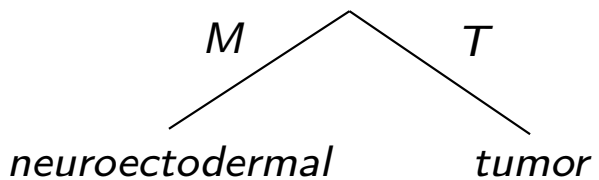
- Analyse syntaxique des groupes nominaux pour en déduire des termes candidats

1. Identification des termes candidats décrits par des patrons d'analyse syntaxique minimaux

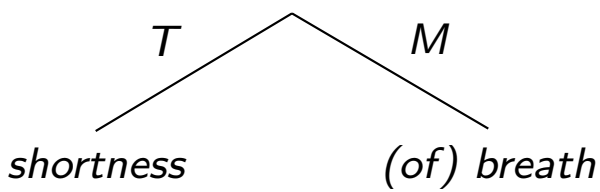


($< T >$: tête du syntagme, $< M >$: modifeur de la tête)

neuroectodermal tumor $\rightarrow$ **(neuroectodermal**$< M >$ **tumor**$< T >$**)**



shortness of breath $\rightarrow$ **shortness**$< T >$ **of breath**$< M >$



YATEA (4)

2. Exploitation des termes candidats analysés précédemment pour analyser les groupes nominaux récursivement

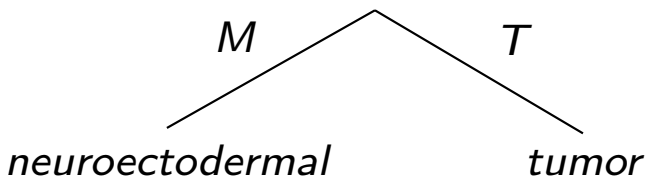
Exemple : **primitive neuroectodermal tumor**

YATEA (4)

2. Exploitation des termes candidats analysés précédemment pour analyser les groupes nominaux récursivement

Exemple : **primitive neuroectodermal tumor**

Exploitation du terme **neuroectodermal tumor** déjà analysé



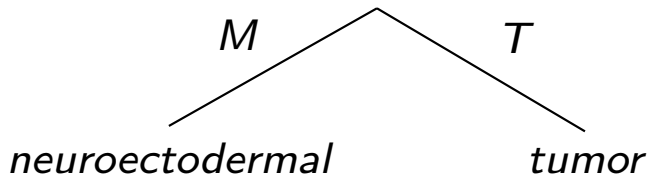
YATEA (4)

2. Exploitation des termes candidats analysés précédemment pour analyser les groupes nominaux récursivement

Exemple : **primitive neuroectodermal tumor**

Exploitation du terme **neuroectodermal tumor** déjà analysé

primitive



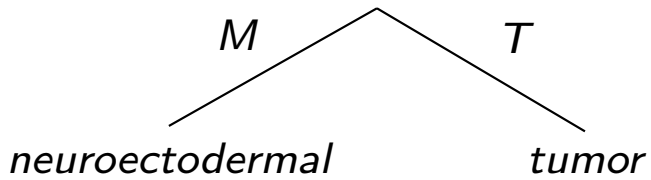
YATEA (4)

2. Exploitation des termes candidats analysés précédemment pour analyser les groupes nominaux récursivement

Exemple : **primitive neuroectodermal tumor**

Exploitation du terme **neuroectodermal tumor** déjà analysé

primitive



Simplification temporaire : primitive$_{JJ}$ tumor$_{NN}$

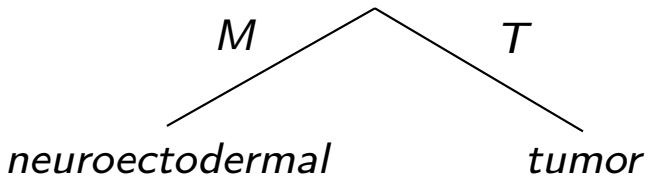
YATEA (4)

2. Exploitation des termes candidats analysés précédemment pour analyser les groupes nominaux récursivement

Exemple : **primitive neuroectodermal tumor**

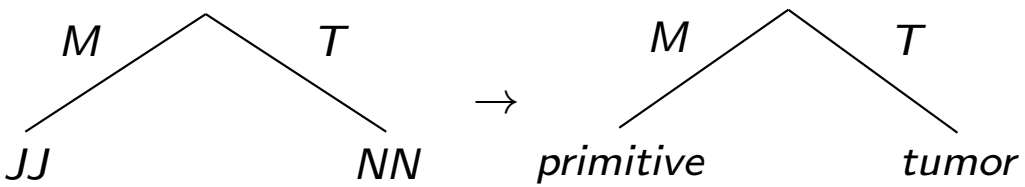
Exploitation du terme **neuroectodermal tumor** déjà analysé

primitive



Simplification temporaire : primitive$_{JJ}$ tumor$_{NN}$

Application du patron :



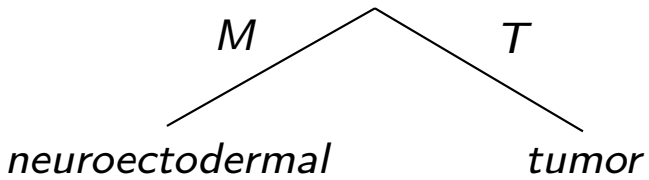
YATEA (4)

2. Exploitation des termes candidats analysés précédemment pour analyser les groupes nominaux récursivement

Exemple : **primitive neuroectodermal tumor**

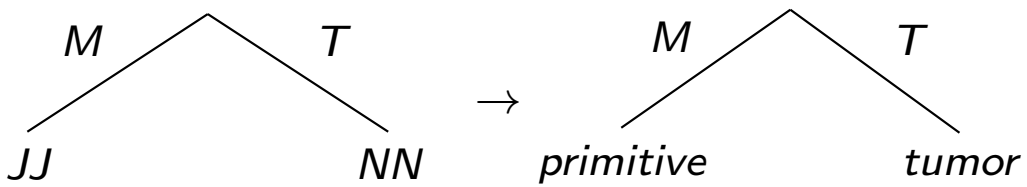
Exploitation du terme **neuroectodermal tumor** déjà analysé

primitive

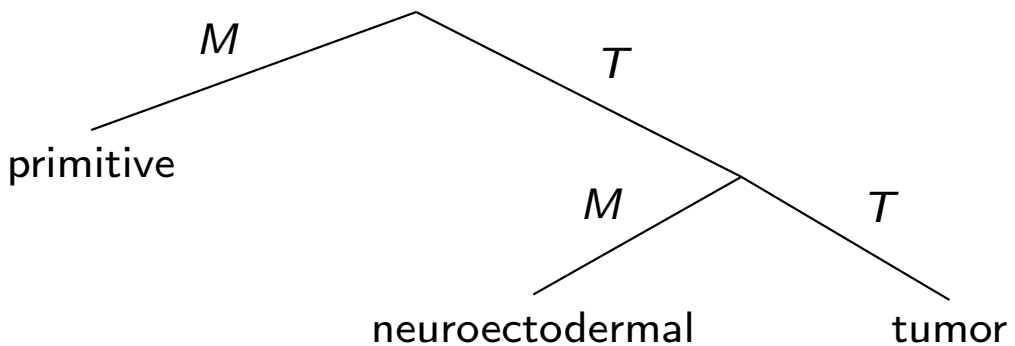


Simplification temporaire : primitive$_{JJ}$ tumor$_{NN}$

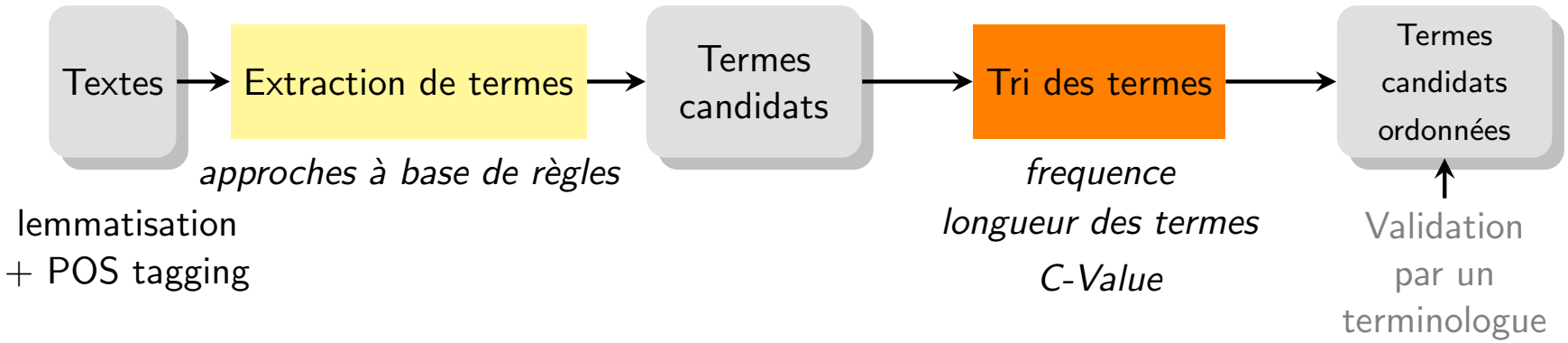
Application du patron :



Redéploiement :



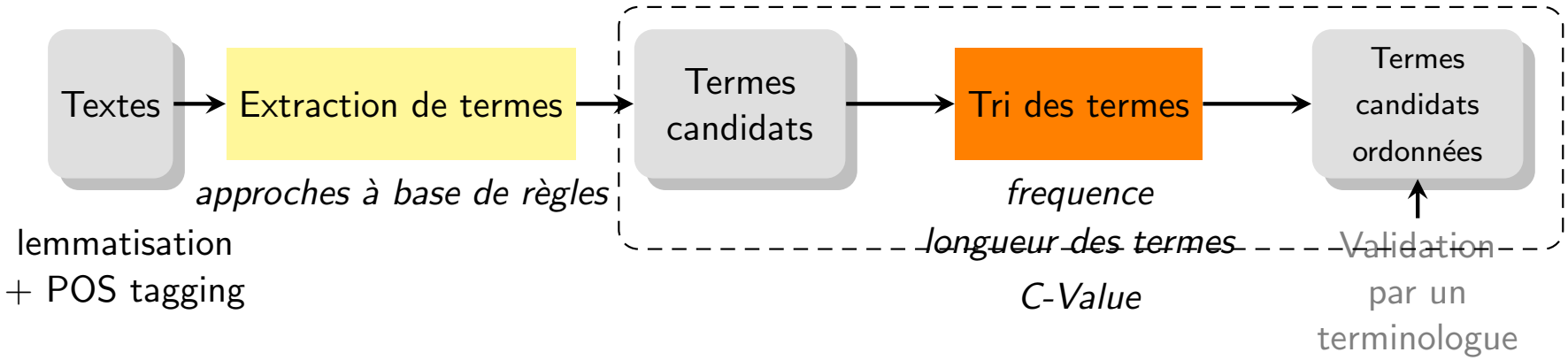
Ordonnancement/Filtrage des termes



- Difficultés:
  - identifier la caractère terminologique des syntagmes extraits
  - ordonner les termes candidats pour identifier les termes du domaine

$\rightarrow$ Définition de métriques pour le tri des termes candidats

Ordonnancement/Filtrage des termes



- Difficultés:
  - identifier la caractère terminologique des syntagmes extraits
  - ordonner les termes candidats pour identifier les termes du domaine

$\rightarrow$ Définition de métriques pour le tri des termes candidats

Métriques pour le tri des termes extraits

- Fréquence : *métrique la plus communément considérée*
  impact variable: dégradation du rappel (beaucoup de termes avec une occurrence) ou la précision

  [Justeson & Katz, 1995, Frantzi *et al.*, 2000, Dowdall *et al.*, 2002]

Métriques pour le tri des termes extraits

- Fréquence : *métrique la plus communément considérée*
  impact variable: dégradation du rappel (beaucoup de termes avec une occurrence) ou la précision
  [Justeson & Katz, 1995, Frantzi *et al.*, 2000, Dowdall *et al.*, 2002]
- Longueur des termes : *les termes longs sont moins important*
  Augmentation légère de la précision quand combinée à la fréquence: les termes simples ou les termes complexes courts sont préférés [Drouin, 2002]

Métriques pour le tri des termes extraits

- Fréquence : *métrique la plus communément considérée*
  impact variable: dégradation du rappel (beaucoup de termes avec une occurrence) ou la précision
  [Justeson & Katz, 1995, Frantzi *et al.*, 2000, Dowdall *et al.*, 2002]
- Longueur des termes : *les termes longs sont moins important*
  Augmentation légère de la précision quand combinée à la fréquence: les termes simples ou les termes complexes courts sont préférés [Drouin, 2002]
- C-Value: *Termes complexes longs qui ne sont pas inclus dans d'autres termes sont préférés* [Frantzi *et al.*, 1997, Frantzi *et al.*, 2000]

$$C-Value_1(t) = \begin{cases} log_2(|t|+1) \cdot f(t) \text{ si } t \text{ n'est pas inclus dans un terme} \\ log_2(|t|+1) \cdot (f(t) - \frac{1}{P(T_t)} \sum_{t' \in T_t} f(t')) \text{ sinon} \end{cases}$$

Amélioration mitigée : précision augmente de 31% pour les termes inclus dans d'autres termes, mais seulement 1% pour tous les termes

Métriques pour le tri des termes extraits

- Fréquence : *métrique la plus communément considérée*
  impact variable: dégradation du rappel (beaucoup de termes avec une occurrence) ou la précision
  [Justeson & Katz, 1995, Frantzi *et al.*, 2000, Dowdall *et al.*, 2002]
- Longueur des termes : *les termes longs sont moins important*
  Augmentation légère de la précision quand combinée à la fréquence: les termes simples ou les termes complexes courts sont préférés [Drouin, 2002]
- C-Value: *Termes complexes longs qui ne sont pas inclus dans d'autres termes sont préférés* [Frantzi *et al.*, 1997, Frantzi *et al.*, 2000]

$$C-Value_1(t) = \begin{cases} log_2(|t|+1) \cdot f(t) \text{ si } t \text{ n'est pas inclus dans un terme} \\ log_2(|t|+1) \cdot (f(t) - \frac{1}{P(T_t)} \sum_{t' \in T_t} f(t')) \text{ sinon} \end{cases}$$

  Amélioration mitigée : précision augmente de 31% pour les termes inclus dans d'autres termes, mais seulement 1% pour tous les termes

- Variante : *NC-value*, prise en compte des termes en contexte
  [Maynard & Ananiadou, 2000]

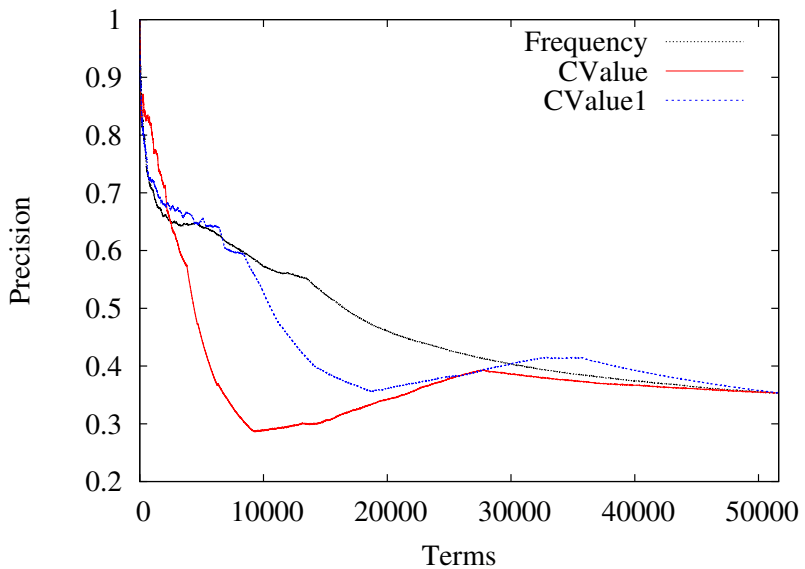
Expériences

- Corpus **Genia** : [Kim *et al.*, 2003]
  - 1 999 résumés Medline (facteurs de transcription dans les cellules humaines)
  - 436 967 mots, 36 607 termes annotés
  - 49 249 termes candidats extraits
- Comparaison avec les termes annotés dans les corpus

Exemples d'évaluation

Comparaison avec les annotations du corpus Genia

- Corpus **Genia** : [Kim *et al.*, 2003]
  - 1 999 résumés Medline (facteurs de transcription dans les cellules humaines)
  - 436 967 mots, 36 607 termes annotés
  - 49 249 termes candidats extraits
- Comparaison avec les termes annotés dans les corpus



Bilan

- Une multitude d'approches utilisant des informations
  - linguistiques
  - ou statistiques
  - ou (plus souvent) les deux
- Des améliorations possibles :
  - Tri des termes candidats pour faciliter le travail du terminologue (les mesures statistiques utilisées ne sont pas toujours convaincantes)
    $\rightarrow$ combinaison de mesures (graphes, regroupement par apprentissage)
  - Association (automatique) de catégories sémantiques aux termes
    $\rightarrow$ vers l'extraction d'événements (beaucoup de travaux sur la reconnaissance d'événements)

Plan

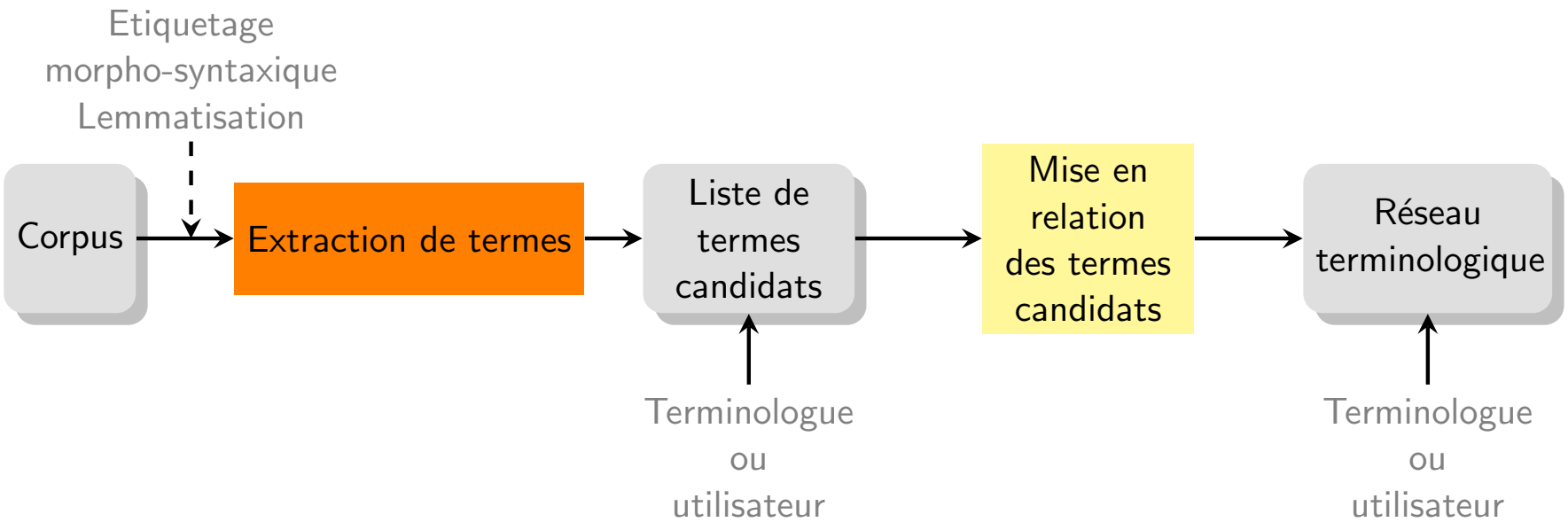
1. Introduction
2. Exemples de terminologies (domaine médical)
3. Constitution de terminologie à partir de corpus
    1. Extraction de termes
    2. *Extraction de relations*

Acquisition de relations

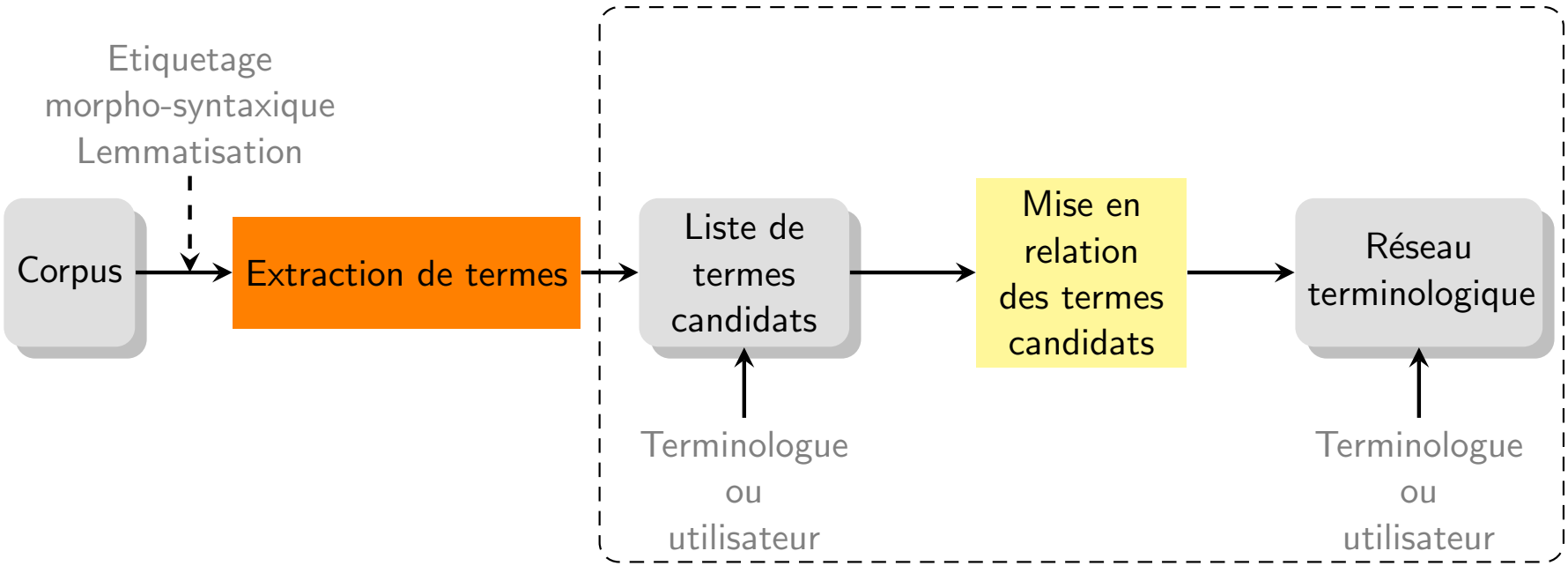
Plan

- Introduction et exemples
- Types de relations sémantiques entre termes
- Approches pour la détection de relations
- Critères de sélection des relations

Acquisition de relations



Acquisition de relations



Structuration de terminologie

Liste de termes : (en général) ensemble de groupes nominaux (termes simples ou complexes)

Objectif: Identifier des relations sémantiques entre les termes
Utilité :

- Normalisation de textes
- Projection de termes sur un texte
- Definition de termes préférés
- Identification de nouveaux termes et de termes obsolètes
- Interoperabilité sémantique (dans les systèmes d'information)
- Enrichissement de terminologie

Deux types d'approches :

- Recherche de variantes de termes (identification de relations)
- Regroupement de termes candidats (construction de classes de termes)

Exemple

Combined action of two transcription factors regulates genes encoding spore coat proteins of Bacillus subtilis.
During sporulation of Bacillus subtilis, spore coat proteins encoded by cot genes are expressed in the mother cell and deposited on the forespore. Transcription of the cotB, cotC, and cotX genes by final sigma(K) RNA polymerase is activated by a small, DNA-binding protein called GerE. The promoter region of each of these genes has two GerE binding sites. 5' deletions that eliminated the more upstream GerE site decreased expression of lacZ fused to cotB and cotX by ...

Exemple

Combined action of two **transcription factors** regulates genes encoding **spore coat proteins** of Bacillus subtilis.
During **sporulation of Bacillus subtilis**, **spore coat proteins** encoded by **cot genes** are expressed in the **mother cell** and deposited on the **forespore**. **Transcription of the cotB**, cotC, and **cotX genes** by final **sigma(K) RNA polymerase** is activated by a small, **DNA-binding protein** called **GerE**. The **promoter region** of each of these **genes** has two **GerE binding sites**. 5' deletions that eliminated the more **upstream GerE site** decreased **expression of lacZ** fused to cotB and cotX by ...

Variantes de termes

gene expression $\Longleftrightarrow$ expression of gene
homologous intramolecular recombination $\rightarrow$ homologous recombination
cytotoxic T-cell $\Longleftrightarrow$ cytotoxic T-lymphocyte
NAD catabolism $\Longleftrightarrow$ nicotinamide adenine dinucleotide catabolism
lymphocyte selection $\Longleftrightarrow$ cell survival
cell proliferation and survival $\rightarrow$ cell survival

Terminologies et relations

- Liste de termes “ à plat ”
- Termes reliés avec des liens non-étiquetés
- Terminologie structurée étiquetée
  - Relations hiérarchiques
  - Synonymes
  - Relations transversales

$\Rightarrow$ Le sens du terme est défini dans sa définition

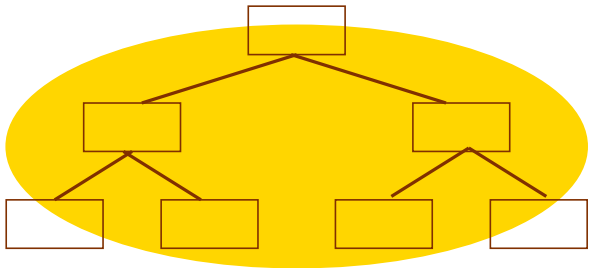
$\Rightarrow$ Le sens du terme est défini par rapport aux autres termes

$\Rightarrow$ Le sens du terme est encodé à travers ses relations avec d'autres termes

Types de relations sémantiques entre termes

- Hyperonymique *vs.* hiérarchiques
- Synonymiques
- Antonymiques
- Transversales
- Associatives

Relations hiérarchiques

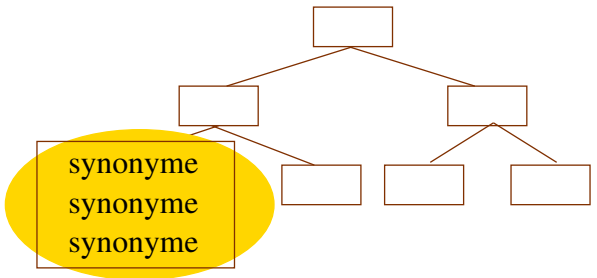


- Relation du générique au spécifique
  relation de subsomption, relation `est-un`
  relation taxinomique, relation hyperonymique (hyponymique)
- Même arbre hiérarchique

`D2-53000` *pneumoconiose, SAI* >
`D2-53400` *pneumopathie liée à l'inhalation de poussière, SAI*

`D5-46000` *maladie de l'appendice, SAI* >
`D5-46100` *appendicite, SAI*

Relations synonymiques



- Relation entre termes équivalents
- Même noeud conceptuel

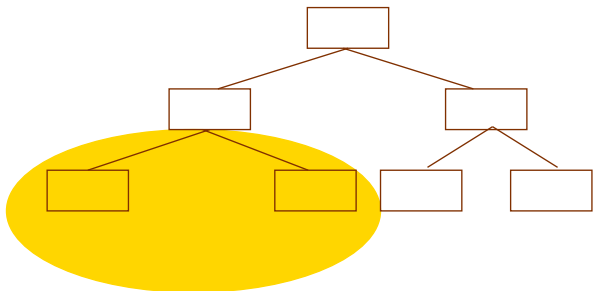
D2-50140 : *pneumonie* ; *pneumopathie inflammatoire*

T-59200 : *appendice vermiculaire* ; *appendico*

D0-10430 : *pemphigoïde, SAI* ; *pemphigus bénin, SAI*

D6-50530 : *déficit en galactose épimérase* ; *galactosémie type III*

Relations antonymiques

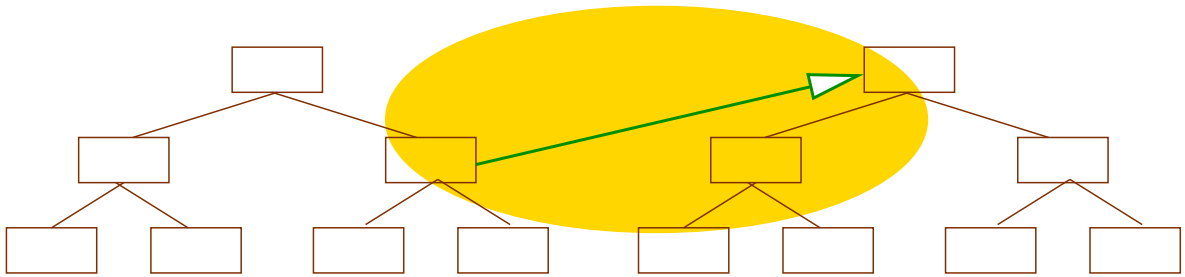


- Relation entre termes opposés
- Entre deux frères non synonymes

D5-45480 : *polypose rectocolique non familiale*
$\Leftarrow \Rightarrow$ D5-45490 : *polypose rectocolique familiale*

D0-00012 : *dermatite aiguë, SAI*
$\Leftarrow \Rightarrow$ D0-00016 : *dermatite chronique, SAI* ;

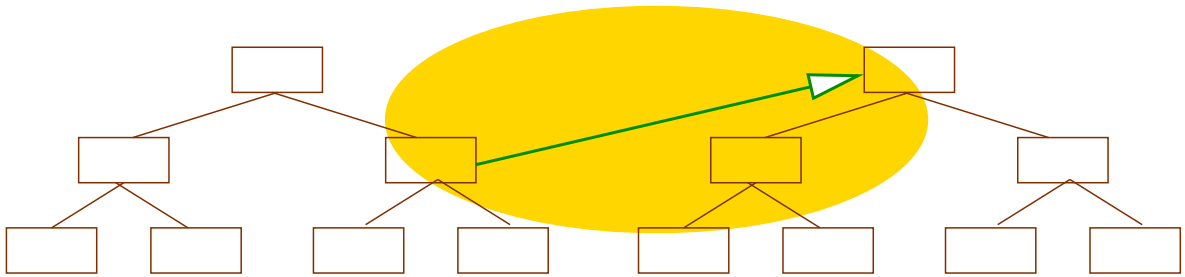
Relations transversales



- Différents arbres hiérarchiques
- Relations thématiques, relations domaniales

- D2-50140 *pneumopathie inflammatoire, SAI* $\mapsto$
  T-28000 *poumon, SAI*
- P1-57450 *appendicectomie, SAI* $\mapsto$
  T-59200 *appendice vermiculaire, SAI*

Relations associatives



- Différents arbres hiérarchiques
- Relation `voir-aussi`, relations non spécifiées

- `A-04242` *lunettes*
  `DA-70000` *maladie de l'oeil, SAI*
  `T-AA610` *rétine, SAI*

Approches pour la détection de relations

Approches en contexte :

- Dépendances syntaxiques
- Patrons lexico-syntaxiques
- Règles d'association
- Approche distributionnelle
- Apprentissage supervisé

Approches hors contexte :
(prise en compte de la structure des termes)

- Règles de substitution
- Inclusion lexicale

Dépendances syntaxiques

- Hypothèse : les termes qui sont en dépendance syntaxique dans une phrase ont une relation sémantique entre eux
- Identification de tout type de relations mais sans connaître le type de la relation
  surtout des relations hiérarchiques et transversales
- Dans la même phrase
- Qualité des résultats dépendant des performances de l'analyseur syntaxique

La probabilité d'infection liée au cathéter est plus faible avec la voie sous clavière

Inclusion lexicale

- Objectif : identification de relations hiérarchiques entre termes
- Hypothèse d'inclusion lexicale
  si un terme est inclu dans un autre terme, il existe une relation d'hyperonymie entre eux
- Inclusion :
  - au niveau de la chaîne de caractères (une chaîne/un mot est inclu(e) dans un groupe de mots)
  - utilisation de relations syntaxiques (tête syntaxique d'un terme)

Inclusion lexicale

- Différents types de relations
  - Hyperonymie :

    venous oxygen saturation `is-a` *oxygen saturation*
    respiratory tract neoplasm `is-a` *neoplasm*
    (Tête) *poumon eosinophile* `est-une` *poumon*

    → Exception : *Generalized Lie theory* / *Lie theory*
  - Relations transversales :

    differential white blood cell count – *white blood*
    hepatite b – *virus hepatite b*

Patrons lexico-syntaxiques

- Hypothèse : Relations sémantiques sont exprimées dans les textes
- Utilisation d'informations lexicales et syntaxiques
  - verbes, noms déverbaux
  - *localisation*, *production*, ...
- Détection de schémas lexico-syntaxiques
  NP e.g. NP_LIST
- Différents types de relations
  synonymie, hiérarchie, transversales, etc.

Projection de patrons lexico-syntaxiques

- Recherche de relations hiérarchiques entre termes
  - à l'aide de patrons construits manuellement (Hearst 1992)
  - en utilisation d'une mesure de similarité ou la programmation logique inductive (Morin 1999, Claveau&L'Homme2004)
- Définition de patrons lexico-syntaxiques à partir de relations d'hyperonymie fournies par un thesaurus
- Exemples :

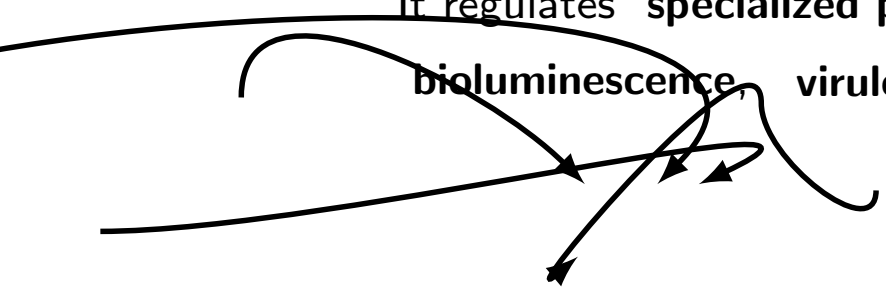
  (Hearst 1992) : NP such as NP, NP, NP, ...

  It regulates **specialized processes** such as **genetic competence** , **bioluminescence**, **virulence**, and **sporulation**.

Projection de patrons lexico-syntaxiques

- Recherche de relations hiérarchiques entre termes
  - à l'aide de patrons construits manuellement (Hearst 1992)
  - en utilisation d'une mesure de similarité ou la programmation logique inductive (Morin 1999, Claveau&L'Homme2004)
- Définition de patrons lexico-syntaxiques à partir de relations d'hyperonymie fournies par un thesaurus
- Exemples :

  (Hearst 1992) : NP such as NP, NP, NP, ...



Projection de patrons lexico-syntaxiques

Identification automatique des patrons (Morin 1999)

des cations tels que le sodium, le potassium, le calcium et le magnésium peuvent être dosés par une méthode de routine

→ SN tel que SN, SN, SN, ...

Acquisition du lien *cultures exigeantes en soufre* / *colza* dans

> Devant le développement de cultures exigeantes en soufres telles que le colza, les agronomes s'intéressent à nouveau au cycle du soufre dans le sol.

⟶ Spécialisation des patrons par apprentisssage

Projection de patrons lexico-syntaxiques

Exemple 2

ResD, when it undergoes ResE-dependent phosphorylation, is thought to activate transcriptionally anaerobically induced genes such as fnr, hmp and nasD.

→ NP such as NP, NP, NP, ...

Acquisition des relations :
specialized processes / *genetic competence*
specialized processes / *bioluminescence*
specialized processes / *virulence*
specialized processes / *sporulation*

dans la phrase :

> It regulates specialized processes such as genetic competence, bioluminescence, virulence, and sporulation.

Patrons lexico-syntaxiques

Relation de synonymie

- NP (également appelé NP_LIST)
  NP , également appelé NP_LIST
  - *Hypoaldostéronisme génétique (également appelé hyperplasie surrénale congénitale)*
  - *Eclampsie, prééclampsie, également appelée hypertension induite par la grossesse.*
  - *Une maladie des membranes basales minces également appelée "hématurie familiale bénigne" au cours de laquelle l'hématurie macroscopique est inhabituelle ...*
- NP W* aussi connu comme W* NP
  - *L'exotoxine VT-1 est aussi connue comme étant une toxine du type Shiga*

Patrons lexico-syntaxiques

Relation hiérarchique

- NP est un NP
  - *EPREX® est un médicament stimulant la formation d'hématies*
  - *Lorsque cet effet secondaire survient, il est observé avec les autres produits de la classe ; si elle est invalidante, la toux est un motif d'arrêt du traitement.*
  - *L'hypertension artérielle est une entité physiopathologique complexe, dont la définition manométrique n'a rien à voir avec un mécanisme physiopathologique : la définition de l'hypertension artérielle est opérationnelle, pratique et arbitraire.*

Patrons lexico-syntaxiques

Relation hiérarchique

- NP est un NP
  - *La nutrition parentérale (NP) est une nutrition passive et ses deux écueils métaboliques sont le risque d'excès d'apport protéino-énergétique et le défaut d'apport en minéraux et en micronutriments (oligo-éléments et vitamines)*
  - *la NP elle même en est un facteur de risque*

Exemples de patrons (1)

(relation d'hyperonymie)

En français (hyperonymie) :

- `{deux|trois...|2|3|4...} SN ( LISTE )`
- `{certain|quelque|de autre...} SN ( LISTE )`
- `{deux|trois...|2|3|4...} SN : LISTE`
- `{certain|quelque|de autre...} SN : LISTE`
- `{de autre}? SN tel que LISTE`
- `SN, particulierement SN`
- `{de autre}? SN comme LISTE`
- `SN tel LISTE`
- `SN {et|ou} de autre SN`
- `SN et notamment SN`
- `chez SN, SN`

Exemples de patrons (2)

(relation d'hyperonymie)

En anglais :

- `SN e.g. LISTE`
- `SN ( e.g. LISTE )`
- `SN called SN`
- `SN known as SN`
- `SN such as LISTE`

Patrons lexico-syntaxiques

Relations transversales

- localisation (NP est localisé (dans|sur|à) NP)
  - *Le gène mutant responsable du CNF est localisé sur le chromosome 19q13.1.*
  - *L'infection urinaire se caractérise par l'invasion de germes pathogènes dans les urines et les tissus de l'appareil urinaire. Lorsqu'elle atteint le parenchyme rénal, on parle d'infection urinaire haute ou de pyélonéphrite. Lorsqu'elle est localisée à la vessie ou à l'urèthre, il s'agit d'infection urinaire basse, de gravité immédiate moins marquée, mais dont le risque principal est la persistance et surtout la récidive.*

Patrons lexico-syntaxiques

Relations transversales

- localisation (NP se trouve à NP
  NP où se trouve à NP_LIST)
  - *La source principale d'activation du système rénine-angiotensine se trouve au niveau du rein et la majeure partie de l'action des IEC est expliquée par son action sur ce système rénal*
  - *le cortex où se trouvent tous les glomérules et la médullaire dont l'extrémité interne ou papille se projette dans la cavité excrétrice*

Règles d'association

- À partir de la cooccurrence de termes dans un même document
- $Terme_1, Terme_2 \Rightarrow Terme_A, Terme_B, ...$
- Si les termes $Terme_1, Terme_2$ apparaissent dans un document, alors les termes $Terme_A, Terme_B$ y apparaissent également
- Si les termes $Terme_1$, $Terme_2$ et $Terme_A$, $Terme_B$, ... apparaissent dans les mêmes documents, alors il existe une relation sémantique entre ces termes
- Pondération des règles avec un indice de confiance " fidélité " des termes dans les documents analysés
- Plusieurs types de relations
- Typage de relations manuel

Règles d'association

Typage manuel

- Hyperonymie
  histamine `est-un` *biogenic amine*
- Co-hyponymie
  spermidine `est-un-frère` *putrescine*
- Relations transversales
  acids `se-transforment-en` *esters*
  silica `utilisé-pour` *chromatography*

Approche distributionnelle

- Basée sur le travail de Zellig Harris
- Calcul du sens des unités lexicales par rapprochement contextuel
- Regroupement des termes partageant suffisamment de contextes (Habert&al 1996, Curran 2004)
- Exemple :

  Le **yaourt**$_{target}$ contient du *calcium*$_{context}$
  Le **lait**$_{target}$ est *riche*$_{context}$ en *calcium*$_{context}$
  Le **lait**$_{target}$ est un *composant*$_{context}$ du *fromage*$_{context}$
  ...

  Regroupement des termes : **yaourt**, **lait**, **produit laitier**, ...
- Utilisation de relations de dépendance syntaxique : `objet de`, `sujet de`, etc.

Approche distributionnelle

- Définition de classes de termes

  large : small, important, major, great, various, main, different, field, new

  patient : case, group, child, day, treatment, woman

- Mais différents types de relations
- Typage manuel des relations
  - antonymie: *large*, *small*
  - synonymie: *large*, *important*, *great*
  - méronymie: *patient*, *group*
  - hyperonymie: *patient*, *child*, *woman*
  - relations transversales: *patient*, *treatment*

Apprentissage supervisé

- Annotation manuelle du corpus : données de référence
- Apprentissage supervisé :
  - calcul des spécificités contextuelles
  - calcul des propriétés des séquences
- Nombre suffisant d'exemples
  - pour chaque type de relations
- Bonne précision, rappel +/- faible

Règles de substitution

- Relation `voir` de *Le petit Robert*

  calibre / *qualité*, *bon* / *beau*
- Relation de synonymie entre les termes complexes

  bon calibre et *belle qualité*
  (variation sur la tête et l'expansion)
- Relation pertinente dans le domaine de la médecine cardiovasculaire

Règles de substitution

- Relations de synonymie entre termes complexes
- Principe de compositionnalité
- Hypothèse de la propagation compositionnelle de relations de synonymie
- Deux termes complexes sont considérés comme synonymes si leurs composants sont identiques ou synonymes
- Ressources lexicales existantes
  thesaurus, dictionnaires de langue générale

Règles de substitution

Types de substitutions :

1. Variation sur l'expansion :
   - les têtes sont identiques et les expansions sont synonymes
   - *action de protection* / *action de sauvegarde*
2. Variation sur la tête :
   - les têtes sont synonymes et les expansions sont identiques
   - *capacité faible* / *puissance faible*
3. Variation sur la tête et l'expansion :
   - les têtes sont synonymes et les expansions sont synonymes
   - *classement d'équipement* / *classification de matériel*

Faster

Christian Jacquemin (1999), Syntagmatic and paradigmatic representations of terms variation. *Proc of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, p. 341-348

Recherche de termes et de leurs variantes morpho-syntaxiques

- Analyseur syntaxique robuste
- Reconnaissance en corpus de termes contrôlés

$\Longrightarrow$ Identification de variantes de termes

- Un ensemble élaboré de métarègles
  pour le repérage de différents types de variation
- Grammaire Syntagmatique Généralisée (HPSG)
  avec unification des structures de traits

Faster

Types de variantes morpho-syntaxiques :

- Variantes syntaxiques :
  - *mesure de volume et de flux*
  - variante de coordination de *mesure de flux*
- Variantes morpho-syntaxiques :
  - *flux de sève mesurés*, *mesure quotidiennement le flux*
    variantes verbales de *mesure de flux*
  - $\Longleftarrow$ relation morphologique entre *mesure/NOM* et *mesurer/VER*
- Variantes sémantico-syntaxiques :
  - *évaluation du flux*
    variante sémantico-syntaxique de *mesure de flux*
    $\Longleftarrow$ proximité sémantique entre *mesure* et *évaluation*

Faster

- Outil d'indexation automatique contrôlée
- Acquisition massive de candidats termes n'est pas son objectif premier
- Une première liste de termes disponible
- Exploitation de cette liste pour reconnaître des variantes de termes
  - $\rightarrow$ enrichissement de la liste

Propagation compositionnelle de relations

Détection de relations de synonymie (Hamon 2000) (SynoTerm)

- Exploitation d'informations de la langue générale
- Inférence de relations de synonymie entre des termes complexes
  Lien inital : *énergie / puissance*
  $\rightarrow$ *évacuation d'énergie / évacuation de puissance*
- Augmentation du rappel en combinant différentes ressources lexicales

$\longrightarrow$ Augmenter la précision en filtrant les liens inférés en fonction du contexte et des ressources utilisées

Inclusion lexicale

- {*sténose*, *sténose aortique*}
- Termes syntaxiquement complexes
- Information syntaxique
  - analyse syntaxique
  - calcul de la tête syntaxique
- Relation hiérarchique

Bilan

Grande variété de méthodes pour acquérir des relations

- Inclusion lexicale : bonne qualité des résultats
- Patrons lexico-syntaxiques : bonne précision des résultats mais le coût de définition des patrons peut être important
- Règles de substitution : qualité variable en particulier pour l'acquisition de relation de synonymie (plus de risque)

Également: méthodes d'apprentissage pour extraire des relations spécifiques

Quelques perspectives d'amélioration

- Réduire l'intervention de l'utilisateur
- Typer les relations (suivant les méthodes utilisées)
- Prendre en compte le contexte, des informations sémantiques associées aux termes et des indices statistiques
- Combiner plusieurs approches ou plusieurs sources de relations


Aubin, S. & Hamon, T. (2006).
Improving term extraction with terminological resources.
In T. Salakoski, F. Ginter, S. Pyysalo & T. Pahikkala, Eds., *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, number 4139 in LNAI, pp. 380–387: Springer.


Bodenreider, O., Rindflesch, T. C. & Burgun, A. (2002).
Unsupervised, corpus-based method for extending a biomedical terminology.
In *Workshop on Natural Language Processing in the Biomedical Domain (ACL2002)*, pp. 53–60.


Bourigault, D., Fabre, C., Frérot, C., Jacques, M.-P. & Ozdowska, S. (2005).
Syntex, analyseur syntaxique de corpus.
In *Actes de la conférence TALN 2005*, pp. 17–20, Dourdan, France.


Cabré, M. T., Estopà, R. & Vivaldi, J. (2001).
Automatic term detection: a review of current systems.
In *Recent Advances in Computational Terminology*. Amsterdam, Philadelphia: John Benjamins.


Cohen, K. B. & Demner-Fushman, D. (2013).
Biomedical Natural Language Processing.
John Benjamins publishing company.


Daille, B. (1995).
Repérage et extraction de terminologie par une approche mixte statistique et linguistique.
T.A.L., **36**(1-2), 101–118.


Dowdall, J., MichaelHess, Kahusk, N., Kaljurand, K., Koit, M., Rinaldi, F. & KadriVider (2002).
Technical terminology as a critical resource.
In *Proceedings of LREC'2002*.


Drouin, P. (2002).
Acquisition automatique des termes : l'utilisation des pivots lexicaux spécialisés.
PhD thesis, Université de Montréal.

Frantzi, K. T., Ananiadou, S. & Mima, H. (2000).

Automatic recognition of multi-word terms: the C-Value/NC-Value method.
International Journal on Digital Libraries, **3**(2), 115–130.

Frantzi, K. T., Ananiadou, S. & Tsujii, J. (1997).

Automatic term recognition using contextual clues.
In *Proceedings of the Second Workshop on Multilingality in software Industry: The AI Contribution (MULSAIC'97), , 15th International Joint Conference on Artificial Intelligence, IJCAI'97*, pp. 73–79, Nagoya, Japan.

Golik, W., Bossy, R., Ratkovic, Z. & Nédellec, C. (2013).

Improving term extraction with linguistic analysis in the biomedical domain.
In *Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'13)*, Samos, Greece.

Grabar, N. & Hamon, T. (2004).

Les relations dans les terminologies structurées : de la théorie à la pratique.
Revue d'Intelligence Artificielle, **18**(1), 57–85.

Hamon, T., Engström, C. & Silvestrov, S. (2014).

Term ranking adaptation to the domain: genetic algorithm based optimisation of the C-Value.
In Springer, Ed., *Proceedings of PolTAL 2014 – Advances in Natural Language Processing*, volume 8686 of *LNAI*, pp. 71–83.

Jacquemin, C. (1997).

Variation terminologique: Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus.
Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes.

Justeson, J. S. & Katz, S. M. (1995).

Principled disambiguation : Discriminating adjective sense with modified nouns.
Computational Linguistics, **21**(1), 1–27.

Kageura, K. & Umino, B. (1996).

Methods of automatic term recognition - a review.

Terminology, **3**(2), 259–89.

Kim, J.-D., Ohta, T., Teteisi, Y. & Tsujii, J. (2003).

Genia corpus - a semantically annotated corpus for bio-textmining.
Bioinformatics, **19**(1), 180–182.
Oxford University Press.

Maynard, D. & Ananiadou, S. (2000).

Identifying terms by their family and friends.
In *Proceedings of COLING 2000*, pp. 530–536, Saarbrucken, Germany.

McCray, A. T., Browne, A. C. & Bodenreider, O. (2002).

The lexical properties of the gene ontology (GO).
In *Proceedings of the AMIA 2002 Annual Symposium*, pp. 504–508.

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C. & Hurdle, J. F. (2008).

Extracting information from textual documents in the electronic health record: a review of recent research.
IMIA Yearbook of Medical Informatics, **42**(5), 923–936.

Nenadic, G., Ananiadou, S. & McNaught, J. (2004).

Enhancing automatic term recognition through recognition of variation.
In *Proceedings of Coling 2004*, pp. 604–610, Geneva, Switzerland: COLING.

Pazienza, M. T., Pennacchiotti, M. & Zanzotto, F. (2005).

Terminology extraction: An analysis of linguistic and statistical approaches.
In S. Sirmakessis, Ed., *Knowledge Mining*, volume 185 of *Studies in Fuzziness and Soft Computing*, pp. 255–279. Springer Berlin Heidelberg.

Spasić, I., Greenwood, M., Preece, A., Francis, N. & Elwyn, G. (2013).

Flexiterm: a flexible term recognition method.
Journal of Biomedical Semantics, **4**, 27.