Computational Terminology

Thierry Hamon
thierry.hamon@univ-paris13.fr

LIM&BIO – EA 3969
Université Paris 13 - UFR Léonard de Vinci
74, rue Marcel Cachin, F-93017 Bobigny cedex
http://www-limbio.smbh.univ-paris13.fr/membres/hamon/

Lim&Bio

April 2012
Introduction

Examples of terminologies (from biomedical area)

Building a terminology from a corpus
  1. Term extraction/recognition
  2. Relations acquisition/extraction
**Introduction**

- Example of terminology purpose
- What is a/the terminology?
- Use of terminologies
- Related professions
- Terminology and corpus
- Documentation, storage
- Creating corpora for terminology building
Example of terminology purpose

Identification of relevant information for a given task:

- Information retrieval: identification of the specific expressions (terms) of each document, adding relations between them, or grouping them in 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.
Example of terminology purpose

Identification of relevant information for a given task:

- Information retrieval: identification of the **specific expressions** (terms) of each document, adding relations between them, or grouping them in 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**.
Example of terminology purpose

Identification of relevant information for a given task:

- Information retrieval: identification of the specific expressions (terms) of each document, adding relations between them, or grouping them in 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.
Example of terminology purpose

Identification of relevant information for a given task:

- Information retrieval: identification of the specific expressions (terms) of each document, *adding relations between them*, or grouping them in 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.
Example of terminology purpose

Identification of relevant information for a given task:

- Information retrieval: identification of the specific expressions (terms) of each document, adding relations between them, or grouping them in 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.
Example of terminology purpose

Identification of relevant information for a given task:

- Information retrieval: identification of the specific expressions (terms) of each document, adding relations between them, or grouping them in 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.
Example of terminology purpose

Identification of relevant information for a given task:

- Information retrieval: identification of the specific expressions (terms) of each document, adding relations between them, or grouping them in 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.
Example of terminology purpose

- Information extraction: fill in a questionnaire to describe an *event*

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.
Example of terminology purpose

Information extraction: fill in a questionnaire to describe an *event*

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.
Example of terminology purpose

Information extraction: fill in a questionnaire to describe an event

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.
What is a/the terminology?

- Science, methods, content
- General purpose Language vs. Specialised Language
- Specialised area: *Technical and Scientific domain*
  - Specificity of the topic
described objects: specific reference (content, meaning)
  - Socio-cultural specificity
constrained practices: know-how, knowledge...
In specialised texts, terminologies are the main part of the lexicon, the remaining part is general and fundamental vocabulary.

*Alain Rey, La terminologie. Noms et notions (translation)*

Medical domain:
- about 400,000 “medical words”
- about 10,000 “common words”
Terminology vs. lexicon, etc.

Lexicon

- Learning, translation
- Mono or multilingual
- Lexical entry
- Related linguistic information
  - phonetic, morphological, syntactic information
- Automatic processing (machine-readable lexicon)
Terminology vs. lexicon, etc.

Lexicographic dictionaries, language dictionaries

(Collins, Le Robert, ...)

- Monolingual
- General language
- Dictionary entry
- Related linguistic information
- Definition
- Examples of use
- Disambiguation
Terminology vs. lexicon, etc.

Encyclopedias – Wikipedia

- Role: didactics
- Entry
- Definition
- Not dedicated to a specific topic
- Common nouns, proper nouns
- Illustrations, pictures
Terminology vs. lexicon, etc.

Terminologies

(Terminological data sheet)

- Given domain
- Meet a need or a function:
  - normalisation, description, transmission of knowledge
- Entries: terms
- (definition, example)
- General term, specific terms (hierarchical relations)
- Other relations (synonymy, etc.)
- Identifier (illustrations)
Terminology vs. ontology

Terminology:
- Terms represent a conceptual system in a particular domain
- Association of linguistic information (linguistic entities)
- No formal organisation of the terms (potential problems: cyclicity, redundancy)
- Designed for humans but contains useful information for an automatic processing (NLP)

Ontology:
- Formal description of concepts and relations of a particular domain
- Association of properties to the concepts
- Designed for automatic processing (inference, computing and formal calculus)

CAVEAT: the both resources are complementary but be careful to not merge them

A terminology can be useful to populate an ontology
Different kinds of terminologies

- Normative terminology
- Descriptive Terminology
- Terminology as a knowledge repository
Observation:
The language of the arts is very imperfect because:

- few proper nouns
- there is a lot of synonyms

For instance, a hammer: as many denomination as arts

More recently (beginning of the 20th century), Vienna school, General theory of the terminology (Wüster 1981):

- Universality of the notions
- Helpful for the communication and translation
Normative terminology

Most common situation: two different denominations

- Regional/local variation
  - In the same language
    - recommandations (Fr – recommendation) vs. ligne directrice (Ca – guideline)
  - Different language
    - maladie de Weber-Christian (Fr) vs. Pfeifer-Weber-Christian-Syndrom/Pfeifer-Weber-Christian-Krankheit (De)

- Sale strategy – technical differenciation
  to identify a difference between similar products or specificity of a brand comparing to others

(in French) airbag vs. coussin de sécurité (safety cushion) vs. coussin gonflable (de sécurité) ((safety) inflatable cushion)
heat pump vs. thermal pump vs. thermopump
Normative terminology

Most current situation: different denominations

- **Speaker**
  - Idiolect
    - *stenosis of the aorta* vs. *aortic stenosis*
  - Specialisation, teaching
    - *myocardial infarction* vs. *heart attack*
    - *rhagade* vs. *chapped skin/crack*

- **Diachronic variation**
  - *oculist* (1503 in French) vs. *ophtalmologue* (1840 in French)
  - vs. *ophtalmologist*
  - Latin root, Greek root
Normative terminology

Application contexts:

- Companies, international organisation, etc.
- Translators, writers
- Technical and official documents
- Designing of the reference terms
- Controlled language
  → Recording of normalised terms
⇒ “ideal Language”
Descriptive terminology

- Automatic processing
- Guarantee the access to the document content
  → need to record the current state of the language in a specialised area
- Record the technological development
  *cogeneration of electricity and heat*
  
  ⇒ “Unrestricted language”
Terminology as a knowledge repository

- Companies
- To face the employee retirement
- Change of employees
- Transfer of knowledge, employee education

→ need to record the current state of the art, of the knowledge

⇒ Experts
⇒ Databases, Knowledge Bases
Users

- Students
- Laymen
- Translators, Technical writers
- Librarians
- Persons in charge of normalisation
- Lexicographers, teachers
- Terminologists
- Knowledge engineers
- NLP
Terminology
Building strategies

- Expert interviews
- Specialised dictionary, existing terminologies
- Corpus exploration
Plan

1. Introduction
2. Examples of terminologies (from biomedical domain)
3. Building a terminology from a corpus
   1. Term extraction/recognition
   2. Relations acquisition/extraction
Examples of terminologies

biomedical domain

- MeSH, document indexing and information retrieval
- ICD, classification of the diseases
- SNOMED, coding the patient records (Electronic Health/clinical Records – EHR/ECR)
- MedDRA, adverse drug effects
- UMLS, unification of medical terminologies
MeSH

document indexing and information retrieval

MeSH, Medical Subject Headings

- document indexing and information retrieval

en: 313,372 concepts, 737,164 labels, 16 types of relation, 879,884 relations

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

se: 25,588 concepts, 25,569 labels, 6 types of relation, 61,036 relations

- synonyms, hierarchical relations and association relations (see-also)

# MeSH, document indexing and information retrieval

## Excerpt from the English version

<table>
<thead>
<tr>
<th>Code</th>
<th>Term</th>
<th>MeSH Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>A01</td>
<td>Body Regions</td>
<td>D001829</td>
</tr>
<tr>
<td>A01.835</td>
<td>Skin</td>
<td>D012867</td>
</tr>
<tr>
<td>A01.835.180</td>
<td>Dermis</td>
<td>D020405</td>
</tr>
<tr>
<td>A01.835.180.360</td>
<td>Hair Follicle</td>
<td>D018859</td>
</tr>
<tr>
<td>A01.835.180.800</td>
<td>Sebaceous Glands</td>
<td>D012627</td>
</tr>
<tr>
<td>A01.835.180.830</td>
<td>Sweat Glands</td>
<td>D013545</td>
</tr>
<tr>
<td>A01.835.180.830.040</td>
<td>Apocrine Glands</td>
<td>D001050</td>
</tr>
<tr>
<td>A01.835.180.830.280</td>
<td>Eccrine Glands</td>
<td>D004439</td>
</tr>
<tr>
<td>A01.835.250</td>
<td>Epidermis</td>
<td>D004817</td>
</tr>
<tr>
<td>A01.835.288</td>
<td>Animal Fur</td>
<td>D006197</td>
</tr>
<tr>
<td>A01.835.288.296</td>
<td>Eyebrows</td>
<td>D005138</td>
</tr>
<tr>
<td>A01.835.288.421</td>
<td>Eyelashes</td>
<td>D005140</td>
</tr>
</tbody>
</table>
# MeSH, document indexing and information retrieval

Excerpt from the Swedish version

<table>
<thead>
<tr>
<th>A01</th>
<th>Kroppsdelar</th>
<th>D001829</th>
</tr>
</thead>
<tbody>
<tr>
<td>A01.835</td>
<td>Hud</td>
<td>D012867</td>
</tr>
<tr>
<td>A01.835.180</td>
<td>Läderhud</td>
<td>D020405</td>
</tr>
<tr>
<td>A01.835.180.360</td>
<td>Hårsäck</td>
<td>D018859</td>
</tr>
<tr>
<td>A01.835.180.800</td>
<td>Talgkörtlar</td>
<td>D012627</td>
</tr>
<tr>
<td>A01.835.180.830</td>
<td>Svettkörtlar</td>
<td>D013545</td>
</tr>
<tr>
<td>A01.835.180.830.040</td>
<td>Apokrina körtlar</td>
<td>D001050</td>
</tr>
<tr>
<td>A01.835.180.830.280</td>
<td>Eckrina körtlar</td>
<td>D004439</td>
</tr>
<tr>
<td>A01.835.250</td>
<td>Överhud</td>
<td>D004817</td>
</tr>
<tr>
<td>A01.835.288</td>
<td>Hår</td>
<td>D006197</td>
</tr>
<tr>
<td>A01.835.288.296</td>
<td>Ögonbryn</td>
<td>D005138</td>
</tr>
<tr>
<td>A01.835.288.421</td>
<td>Ögonfransar</td>
<td>D005140</td>
</tr>
</tbody>
</table>
MeSH, document indexing and information retrieval

Excerpt from the French version

<table>
<thead>
<tr>
<th>A01</th>
<th>REGION</th>
<th>D001829</th>
</tr>
</thead>
<tbody>
<tr>
<td>A01.835</td>
<td>PEAU</td>
<td>D012867</td>
</tr>
<tr>
<td>A01.835.180</td>
<td>DERME</td>
<td>D020405</td>
</tr>
<tr>
<td>A01.835.180.360</td>
<td>FOLLICULE PILEUX</td>
<td>D018859</td>
</tr>
<tr>
<td>A01.835.180.800</td>
<td>GLANDES SEBACEES</td>
<td>D012627</td>
</tr>
<tr>
<td>A01.835.180.830</td>
<td>GLANDES SUDORIPARES</td>
<td>D013545</td>
</tr>
<tr>
<td>A01.835.180.830.040</td>
<td>GLANDE APOCRINE</td>
<td>D001050</td>
</tr>
<tr>
<td>A01.835.180.830.280</td>
<td>GLANDE ECCRINE</td>
<td>D004439</td>
</tr>
<tr>
<td>A01.835.250</td>
<td>EPIDERME</td>
<td>D004817</td>
</tr>
<tr>
<td>A01.835.288</td>
<td>CHEVEU</td>
<td>D006197</td>
</tr>
<tr>
<td>A01.835.288.296</td>
<td>SOURCIL</td>
<td>D005138</td>
</tr>
<tr>
<td>A01.835.288.421</td>
<td>CIL</td>
<td>D005140</td>
</tr>
</tbody>
</table>
ICD, classification of the diseases

(International classification of the diseases, causes of morbidity and connected health problems)

- WHO (World Health Organisation)
- *indexing patient records for statistical needs (surveys, health alters...*)

**en**: 12,318 classes, 13,505 labels

**fr**: 10,800 classes, 9,412 labels

- included terms (synonyms), hierarchical relations
ICD, classification of the diseases

Excerpt from the English version

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
ICD, classification of the diseases

Excerpt from the French version

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, coding the patient records

SNOMED Nomenclature

(Systematised NOmenclature of the MEditional problems) - Intl

- description of clinical information (patient records)

en: 112,661 concepts, 164,069 labels

fr: 9,098 concepts, 12,554 labels, 11,290 hierarchical relations in French

- synonyms, hierarchical relations (is-a), meronomy relations (part-of) et inter-hierarchical relations
SNOMED, coding the patient records

Terminology with several axis

The concepts are organised hierarchically in eleven semantic axes

- Morphology (M)
- Topography (T)
- Function (F)
- Living Organisms (L)
- Agents and physical activities (A)
- Diagnostics (D)
- etc.
SNOMED, coding the patient records

Excerpt from the English version

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, coding the patient records

Excerpt from the French version

<table>
<thead>
<tr>
<th>Code</th>
<th>Code Type</th>
<th>French Term</th>
<th>SAI Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-12000</td>
<td>01</td>
<td>fracture</td>
<td>SAI</td>
</tr>
<tr>
<td>M-12000</td>
<td>05</td>
<td>fracturé</td>
<td></td>
</tr>
<tr>
<td>M-12010</td>
<td>01</td>
<td>fracture transverse</td>
<td></td>
</tr>
<tr>
<td>M-12020</td>
<td>01</td>
<td>fracture oblique</td>
<td></td>
</tr>
<tr>
<td>M-12200</td>
<td>01</td>
<td>fracture ouverte, SAI</td>
<td></td>
</tr>
<tr>
<td>M-12300</td>
<td>01</td>
<td>fracture non consolidée, SAI</td>
<td></td>
</tr>
<tr>
<td>M-12400</td>
<td>01</td>
<td>fracture avec retard de consolidation, SAI</td>
<td></td>
</tr>
<tr>
<td>M-12500</td>
<td>01</td>
<td>fracture consolidée, SAI</td>
<td></td>
</tr>
<tr>
<td>M-12500</td>
<td>02</td>
<td>fracture guérie, SAI</td>
<td></td>
</tr>
<tr>
<td>M-12520</td>
<td>01</td>
<td>fracture avec consolidation fibreuse</td>
<td></td>
</tr>
<tr>
<td>M-12590</td>
<td>01</td>
<td>pseudarthrose, SAI</td>
<td></td>
</tr>
<tr>
<td>M-12590</td>
<td>02</td>
<td>néarthrose</td>
<td></td>
</tr>
</tbody>
</table>
MedDRA, adverse drug effects

*Medical Dictionary for Drug Regulatory Activities*

Objectives:

- Description of the phases of medication development
- Description of the problems related to the medication (from a regulatory point of view)

Utilised by:

- FDA (USA), database AERS, VAERS
- EMEA (European Medical Evaluation Agency), database Eudrawatch
- PEM (Prescription Event Monitoring), Japan

Used as a standard for information exchange in Europe
MedDRA, adverse drug effects

*Medical Dictionary for Drug Regulatory Activities* termes :
- Adverse drug effects
- Indications
- Signs et symptoms
- Family History
- Laboratory data (exams)
- Surgery interventions

5 hierarchical levels
UMLS, Unified Medical Language System

Unification of biomedical terminologies

- gathering and integrating several biomedical terminologies
- 20 languages
- 158 terminologies
- Problem: Semantic interoperability between the terminologies

en: 2,381,083 concepts, 5,491,897 labels (strings)
se: 26,151 concepts, 26,209 labels (strings)
fr: 82,170 concepts, 156,762 labels (strings)

- synonyms, hierarchical relations, about 100 semantic relations (and some types of relations are missing!)
UMLS, unification of medical terminologies

Components of UMLS:

- Metathesaurus
  Source terminologies

- Semantic Network
  Semantic relations between terms

- SPECIALIST Lexicon
  English lexicon (general language, medical language)

- MetamorphoSys
  Tools to make easier the use of UMLS
  Table exploration, visualisation, term matching, ...
Terminology
Approaches for building

- Expert interviews
- Specialised Dictionary, existing terminologies
  ⇒ Corpus exploration
Based on work in Corpus Linguistics

- No hypothesis on the language
- The needed information is in the corpus
- Specialised languages: Considered as sub-languages (in the sense of the formal language), characterised by:
  - limited lexicon (terms, and synonyms)
  - specific syntactic patterns
- Availability of NLP tools for mining and processing the corpus
How to build a corpus

- Opportunistic collection of texts
- Targeted collection of texts:
  - representativeness:
    - sufficient amount of texts
    - various parameters
      - type of the texts, specialisation, topics, etc.
  - recording of the origin
  - ...

Requirements regarding the corpus

A corpus has to:

- be relevant according to the domain
  selected texts represent those written in the domain
- be relevant according to the application
  selected texts represent those used in the final application
- take into account the capacity of the NLP tools
Relevance according to the domain

- Texts contain knowledge of the domain:
  - specialised texts
  - ...
  - vulgarisation texts
Methods for building a corpus

- Scanner
- Collecting texts from experts
- Use of the company/department documentation
- Exploitation of text database (in biomedical domain: Medline)
- Exploitation of text from Internet (discussion from support groups, blogs, etc.) – but be careful!
How to evaluate the relevance according to the domain?

- Description of the domain:
  - keywords, main terms, descriptors
- “Centrality” of the documents:
  - occurrence of the keywords in the documents
- Weighting the document with vectorial methods \( \Rightarrow \)
  - a document is represented by the descriptors it contains
- Pagerank?
Relevance according to the application

Take into account the end-user application:

- **Objectives**
  Study of linguistic phenomena, grammar, discourse

- **Application**
  Information retrieval, indexing, information extraction, etc.

- **Scope**
  Inside, outside a company, at the national level, etc.

- **Specialisation level**
  General language corpus vs. specialised language corpus vs. non-expert texts

- **Type of corpus**
  Multilingual context: parallel or comparable corpus
How to evaluate the relevance according to the NLP tools/methods?

- Take into account the capacity of the NLP tools
- Choice of the tools according to the texts to analyse:
  - robustness
  - language
  - format
- Monolingual or multilingual tools
Specialised corpus

- Specialised domain
  medicine, cogeneration, telecommunications, etc.
- Characteristics of a specialised language (sub-language) for a given domain
- Lexical and syntactic specificity
- The first step to build
  - terminologies
  - thesaurus, etc.
- Organising the internal structure:
  sub-corpus by topics, origins, etc.
Specialised corpus

- Gather a sufficient amount of texts
- Anonymisation of texts
- Definition of criteria: type of texts (various styles), sub-domains
Plan

1. Introduction
2. Examples of terminologies (from biomedical area)
3. Building a terminology from a corpus
   ⇒ Term extraction/recognition
   1. Relation acquisition/extraction
Term extraction

- Approaches for extraction and recognition of terms from/in corpus
- Tools for extraction
Term definition

- What is a term?
- What is not a term?
Terms or not?

- Acantholytic dermatosis
- Abdominal cramps
- Accidental Release, Chemical
- has chronic debt problems
- Therapy, Acupuncture
- demethylchlordimeform hydrochloride
- Acetic acids
- Syndrome Adams Strokes
- Adams Strokes, syndrome
Terms or not?

- Eleventh thoracic vertebra
- Eighth cranial nerve disease or syndrome
- Drug or chemical induced diabetes mellitus with ophthalmic complications
- Ectopic ACTH secretion causing Cushing’s syndrome
- ATP-dependent polydeoxyribonucleotide 5’-hydroxyl-kinase activity
- 9-((2’S,3’S)-bis(hydroxymethyl)thietan-1’-yl)adenine
- (6-chloro-2,3-dihydro-1H-inden-1-yl)acetic acid
Terms or not?

- Eleventh thoracic vertebra
- Eighth cranial nerve disease or syndrome
- Drug or chemical induced diabetes mellitus with ophthalmic complications
- Ectopic ACTH secretion causing Cushing’s syndrome
- ATP-dependent polydeoxyribonucleotide 5’-hydroxyl-kinase activity
- 9-((2’S,3’S)-bis(hydroxymethyl)thietan-1’-yl)adenine
- (6-chloro-2,3-dihydro-1H-inden-1-yl)acetic acid

There are all terms, all those phrases are issued from a terminology!
A terminology should exclude:

- enunciation marks:
  - personal pronoun
  - possessive adjectives
  - temporal and location adverbs
- “grammatical” words (empty words / stop words)
- verbs (except when they can be considered as nouns)
Where are the terms?

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 ...
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 ...
Process of terminology building

**Extraction**
- Corpus
  - Part-of-speech tagger
- Analyser
  - List of candidate terms
  - Terminologist or user
- Acquisition of relation between candidate terms
  - Lexico-syntactic patterns
  - Lexical inclusion
  - Substitution rules
  - Distributional analysis
  - Terminological resources

**Structuring**
- Terminological network
  - Terminologist or user
Enriching document with linguistic information

- identification of the sentences

The promoter region of each of these genes has two GerE binding sites.
Enriching document with linguistic information

- identification of the sentences, words (+ lemma and part-of-speech category)

The promoter region of each of these genes has two GerE binding sites
Enriching document with linguistic information

- identification of the sentences, words (+ lemma and part-of-speech category, named entities

The promoter region of each of these genes has two GerE binding sites.
Enriching document with linguistic information

- identification of the sentences, words (+ lemma and part-of-speech category, named entities, and terms

The promoter region of each of these genes has two GerE binding sites
Process of terminology building

Beginning

- Split the sentences considering the syntactic boundaries of terminological phrases
  - pronouns, conjugated verbs
  - prepositions
  - (coordination)
  - punctuation

- Identification of connection words
  - *of, of, the, etc.*

- Identification of lexical anchors
  - already known “central/main” words
Process of terminology building

Extraction

- Search of the repeated textual segments in a window of $n$ words
- Application of syntactic patterns typical for noun phrases
- Application of syntactic patterns typical for noun and adjectival phrases
- Application of syntactic patterns of bi-terms
- Search of the repeated textual segments in the neighbourhood of connection words
- Search of the repeated textual segments in the neighbourhood of lexical anchors
Process of terminology building

Improvement

- Maximal phrase decomposition
- Statistical filters
- Lexical filters
- Use of variation rules
- Merging of term variants
Tools for the term extraction

- Termino, first tool
- Acabit, focus on the bi-terms (two word terms)
- Lester, endogenous learning
- XTract, statistical learning
- TermoStat, measures of termhood
- Toolbox Xelda, noun phrases extraction
- YATEA (CPAN), term acquisition and enrichment
Termino

Sophie David et Pierre Plante (1990)

- Building of terminological database
  ⇒ first tool dedicated to the terminology
- Project of Office de la Langue Française (OLF), Quebec, Canada
  ⇒ Quebec linguistic protection Policy
Termino

- First automatic tool
- Acquisition of terms by analysing corpora
- Based on the syntactic annotation of texts
- Identification of structured noun phrases which seem to be relevant candidate terms

Phases:
1. Text preprocessing
2. Terminological analysis and acquisition
3. Interactive building of term database (term bank)
Termino

Terminological analysis and acquisition

- Acquisition of candidate terms: 3 steps
  1. Rule-based morphological analysis
  2. Identification of noun phrases
     - two level noun phrase grammar
     - X-bar theory (syntagmatic rules)
  3. Generation of the candidate terms
     - dependancies between head/main word and modifiers/dependent words
Termino

Generation of the candidate terms

1. For each detected noun:
   - analysis to the right of the list of tagged words
   - until the syntactic boundary is found
   - identification of the potential modifiers

2. Categorisation of modifiers according to the ambiguous attachment
   - \texttt{traitement de texte gratuit}
   - \textit{observation of the stenosis of the aorta}
   - \textit{gratuit} is marked as ambiguous
   - \textit{aorta} is marked as ambiguous

3. For all the processed texts:
   - collect information associated with the head noun
   - disambiguate some attachments, by occurrence comparison
(Daille, 1995)

- Hybrid approach: based on linguistic and statistical analysis
- Focus on bi-terms and their variants
- Extraction of candidate terms from corpora (pos-tagged)

NB: Languages: French, English, Japanese
Acabit

2 steps:

1. Linguistic analysis and term variants grouping:
   - POS-tagged corpus
   - Transductors to identify noun phrases
   - Extraction of candidate terms:
     - N ADJ : station terrienne (Earth station)
     - N₁ PREP N₂ : liaison par satellite (satellite link)
     - N₁ N₂ : diode tunnel (tunnel diode)
   - Splitting of terms into binary candidate terms:
     - réseau de transit à satellite (satellite transit network)
     - → réseau de transit (transit network)
     - → réseau à satellite (satellite network)
2. Statistical filtering:
   - Computing statistical measures to rank the candidate terms
   - Computing scores, metrics and distances regarding the components of the terms (mainly based on frequency)
   - Log-likelihood ratio (Dunning, 1993)
     the best score to keep candidate terms independently to their frequency
Endogeneous learning (without use of domain knowledge)

- Extraction of candidate terms from a POS tagged corpus
- Shallow parsing of the texts
- Identification and analysis of the (maximal) noun phrases
- Structuring the set of candidate terms into a network (syntactic network)

For French texts
Lexter

1. Extraction of the maximal noun phrases

- Identification of syntactic boundaries
  conjugated verbs, pronouns, subordinating conjunctions...
- Extraction of maximal noun phrases
- Endogenous learning on the corpus
- Collection of information on sub-categorisation of the nouns
  and adjectives occurring in corpus
- Resolving ambiguities of prepositional attachments
- Example: noun *pression* (pressure) sub-categorise the
  preposition à (to):
  - *pression à l’aspiration* (suction pressure)
  - *pression au refoulement* (discharge pressure)
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 ...
Combined action of two transcription factors regulate 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.
Lexter

2. Splitting of the maximal noun phrases

- Recursive analysis of the maximal noun phrases into phrase components
- Identification of the syntactic head and modifiers
- *severe stenosis of left main coronary artery*
Lexter

2. Splitting of the maximal noun phrases

*sténose sévère du tronc commun de l’artère coronaire gauche*
Lexter

2. Splitting of the maximal noun phrases

- Endogenous learning of the prepositional dependency (on the corpus)
- Deal with dependency ambiguity within noun phrases
- Candidate terms:
  - maximal noun phrases
  - their constituents
3. Structuration module

- Syntactic network of candidate terms
- Use of the head and modifier relations ...
- Computing a productivity coefficient
density of the network around a candidate term
Lexter

Example of use of Lexter

**Input:**

```
```

**Output:** extraction of candidate terms:

```
→ (sténose sévère) du (tronc commun de l’((artère coronare) gauche))
→ minimum d’injections
```
XTract

(Smadja, 1993)

Use of statistical approaches

- Extraction of collocations
- Extraction of repeated and preferred lexical associations
- Three modules:
  1. Identification of the collocations
  2-3. Use of these collocations
XTract

1. Extraction of binary collocations

- Extraction of associations
- Pair of close words (often cooccurring or related)
  Hypothesis: two words do not appear together by chance
- Take into account a given distance between words

⇒ Extraction of flexible collocations
  *expensive takeover*
XTract

2. Expansion of collocations

- Iterative search of collocations

  → Identification of collocations with more than two words
XTract

3. Tagging the collocations

- Observation: Extraction of various collocations
- Categorisation of the collocations with the parser Cass (Abney, 1990)
- Three main types of collocations:
  - predicative collocations
    - verbe + predicative noun
      - make decision
  - idioms
    - stock market
  - sub-specified phrases
XTract

- Mirrored design of Acabit
- XTract
  1. statistical selection
  2. linguistic filtering
- Acabit
  1. linguistic filtering
  2. statistical selection
YATEA (1)
Yet Another Term ExtrActor
(Aubin et Hamon, 2006)

- Term Extraction from French and English texts
- Parsing terms with syntactic dependencies
- Rejection of unparsable noun phrases
- Developed in the European Project ALVIS

Perl Module on CPAN
http://search.cpan.org/~thhamon/Lingua-YaTeA-0.3/

Description of the extraction and parsing in configuration files

Simplification of the sentence parsing (time gain and result quality improvement)
Identification of chunks thanks to morpho-syntactic information (frontiers - verbs, adverbs, etc.)

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 ...
Identification of chunks thanks to morpho-syntactic information (frontiers - verbs, adverbs, etc.)

Combined action of two transcription factors genes
spore coat proteins of Bacillus subtilis
sporulation of Bacillus subtilis spore coat proteins
cot genes mother cell
forespore Transcription of the cotB
cotX genes final sigma(K) RNA polymerase
DNA-binding protein called GerE promoter
region of each of these genes two GerE binding sites 5’
deletions more upstream GerE site
expression of lacZ cotB cotX
Syntactic analysis of the noun phrases to detect term candidates

1. Identification of term candidates described by parsing patterns
   \( (\text{Head of the noun phrase}, \text{modifier of the head}) \)

   - mother cell \( \rightarrow \) (mother\(M\) cell\(T\))
   - transcription factor \( \rightarrow \) (transcription\(M\) factor\(T\))
   - expression of lacZ \( \rightarrow \) expression\(T\) of lacZ\(M\)
2. Use of the previously parsed term candidates (island of reliability) to parse remaining noun phrases

**cotC** promoter region

Use of the parsed terms \((\text{promoter}<M>\text{region}<T>)\)

\[ \rightarrow (\text{cotC}<M>\text{promoter}<M>\text{region}<T>) \]
Plan

1. Introduction
2. Examples of terminology (from biomedical area)
3. Building a terminology from a corpus
   1. Term extraction/recognition
   2. Relations acquisition/extraction
Relation acquisition

- Types of semantic relations between terms
- Approaches for the acquisition of relations
- Criteria for the selection of relations
Terminology structuring

Term list: noun phrases and multi-word components

Aim: Adding semantic relations between terms

- Text normalisation
- Term matching
- Defining preferred terms
- Identifying new terms and obsolete terms
- Semantic interoperability (information system)
- Enriching a terminology

Two types of approach:

- Searching term variants (identification of relations)
- Collection of term candidates (building class of terms)
Example

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 ...
Example

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 ...
Term variants

gene expression ⇔ expression of gene
homologous intramolecular recombination → homologous recombination
cytotoxic T-cell ⇔ cytotoxic T-lymphocyte
NAD catabolism ⇔ nicotinamide adenine dinucleotide catabolism
lymphocyte selection ⇔ cell survival
cell proliferation and survival → cell survival
Terminologies and relations

- Terms can be related but the relation is not typed
- Structured terminology (with typed relations)
  - Hierarchical relations
  - Synonyms
  - Inter-hierarchical relations

⇒ The meaning of the term is included in its definition
⇒ The meaning of the term is defined according to its neighbour terms
⇒ The meaning of the term is encoded thanks to the relations with the neighbour terms
Types of semantic relations between terms

- Hyperonymy vs. hierarchical relations
- Synonymy
- Antonymy
- Inter-hierarchical
- Associative
Hierarchical relations

- Relation between a general term and a specific term: subsumption relation, *is-a* taxonomic relation, hyperonymic relation (hyponymic)
- Relation in the same hierarchical tree

- D2–53000 *Pneumoconiosis, NOS >*
- D2–53900 *Radiation pneumonitis*
- D5–46000 *Disease of appendix, NOS >*
- D5–46100 *Appendicitis, NOS*

NB: two terms sharing the same general term: siblings
Synonymous relations

- Relation between terms with similar (equivalent) meaning
- Same conceptual node (in a hierarchy)

D2-50140: Pneumonia, NOS; Pneumonitis, NOS
T-59200: Vermiform Appendix; appendix; Appendix structure
D0-10430: Pemphigoid, NOS; bullous pemphigoid
D6-50530: Galactose epimerase deficiency; Galactosemia III: UDPglucose-4-epimerase deficiency
Antonymous relations

- Relation between terms with opposite (contrary) meanings
- Two terms sharing the same semantic feature but with opposite values

C0751399: *Nonfamilial Olivopontocerebellar Atrophy*
⇐⇒ C0751398: *Familial Olivopontocerebellar Atrophy*

C0023473: *Chronic myeloid leukaemia*
⇐⇒ C0023467: *Acute myeloid leukaemia*
Inter-hierarchical relations

- Relation between terms in two different hierarchical trees
- Thematic relations, domain relations, location relations, etc.

- D2-50140 *Pneumonitis, NOS*  \(\leftrightarrow\)  
  T-28000 *Lung, NOS*

- P1-57450 *Appendicectomy, NOS*  \(\leftrightarrow\)  
  T-59200 *Appendix, NOS*
Associative relations

- Relation between terms in two different hierarchical tree
- Relation: see-also, unspecified relations

- A-04242 *Eye glasses*
- DA-70000 *Disease of eye, NOS*
- T-AA610 *Retina, NOS*
Approaches for extracting relations

In-context approaches:
- Syntactic dependencies
- Lexico-syntactic patterns
- Association rules
- Distributional analysis

Out-of-context approaches:
(based on the analysis of the term structure)
- Substitution rules
- Lexical inclusion
Syntactic dependencies

Simple and coarse-grained approach:

- **Hypothesis:** the terms which are related by a syntactic dependency (in a sentence) are linked by a semantic relation.

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

The probability d’infection related to the catheter is more low with the subclavian artery.

---

**Diagram:**

```
P
  +---+     +---+     +---+
  | NP|     | SP|     | VP|
  +---+     +---+     +---+
     | DET  |  | DET  |  | PREP|
     | N    |  | N    |  | N   |
     | PROBABILITE | DE | INFECTION |
     +---+     +---+     +---+
           | SP|     | SP|     | SP|
           | PREP|     | PREP|     | PREP|
           | N   |     | N   |     | N   |
           | PROBABILITE | INFECTION | plus | faible | voie |
           +---+     +---+     +---+     +---+     +---+
               | ADV |  | ADJ |  | DET |  | ADJ |
               | est |  | ADJ |  | N   |  | sous |
               | PLUS |  | FAIBLE |  | VOIE |  | CI |
Syntactic dependencies

- Identification of all kind of relations (but without identification of their type)
  especially inter-hierarchical relations
- In the same sentence
- Depends on the quality of the syntactic parsing
Lexico-syntactic patterns

- Hypothesis: semantic relations are expressed/encoded in the texts
- Based on lexical and syntactic information
  - verbs, deverbal nouns (nominalisation)
  - localisation, production, construct, ...
- Lexico-syntactic patterns can be defined
- Various types of relations
  hyperonymy, inter-hierarchical relations, synonymy, etc.
Hierarchical relations

• is a or such as

  • *ermC is a plasmid gene which specifies resistance to macrolide-lincosamide-streptogramin B antibiotics.*

  • *Inhibition of RNA polymerase activity by gramicidin in consistent with a sigma-specific effect: the antibiotic is a strong inhibitor of transcription of T7 phage DNA, which requires sigma for activity, but it has little effect on transcription of sigma-independent templates, such as poly(dA-dT).poly(dA-dT) and calf thymus DNA.*
Lexico-syntactic patterns

Inter-hierarchical relations

- localisation (is/were localised)
  - All spontaneous suppressor mutations obtained from a secA12 sporulation-defective mutant in Bacillus subtilis were localized in highly conserved membrane-spanning regions of SecY. The expression of early
  - one pole of the bacterium. IcsA, an outer membrane protein, is localized to the old pole of the bacterium and is both necessary and sufficient for actin assembly. IcsA is slowly cleaved from the bacterial surface by the
Lexico-syntactic patterns

Inter-hierarchical relations

- activation (is activated)
  - Transcription of srf is activated by the two-component regulatory system ComPA in response to the peptide pheromone, ComX, which mediates cell density-dependent control.
  - Both RapC and CSF were found to be part of autoregulatory loops that affect transcription from P1, which we show is activated by ComA approximately P. RapC negatively regulates its own expression, presumably due to its ability to inhibit accumulation of ComA approximately P. CSF positively regulates its own expression, presumably due to its ability to inhibit RapC activity.
Lexico-syntactic patterns

Synonymy relation

- Also called
  - The Epstein-Barr Virus (EBV) protein EB2 (also called Mta, SM and BMLF1), is an essential nuclear protein produced during the replicative cycle of EBV.
  - Glucosamine (UDP-N-acetyl)-2-epimerase/N-acetylmannosamine kinase (GNE) myopathy, also called distal myopathy with rimmed vacuoles (DMRV) or hereditary inclusion body myopathy (HIBM), is a rare, progressive autosomal recessive disorder caused by mutations in the GNE gene.

- (Also) known as
  - Two new regulatory genes were identified in our screen: ypuN (also known as rsiX, the anti-sigmaX factor) and ylbF.
  - A mechanism for regulating gene expression at the level of transcription utilizes an antagonist of the sigma transcription factor known as the anti-sigma (anti-sigma) factor.
Use of lexico-syntactic patterns

Identification of hyperonymic relations with lexico-syntactic patterns defined manually (Hearst 1992) by using similarity measure and ILP (Morin 1999, Claveau&L’Homme2004)

- Exploitation of hyperonymy relations issued from a thesaurus to acquire lexico-syntactic patterns

- Example:
  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 of the relations:
- specialized processes / genetic competence
- specialized processes / bioluminescence
- specialized processes / virulence
- specialized processes / sporulation

in the sentence:

It regulates specialized processes such as genetic competence, bioluminescence, virulence, and sporulation.
Examples of patterns (2)
(hyperonymy relations - Hearst1992)

- NP e.g. LISTE
- NP ( e.g. LISTE )
- NP called NP
- NP known as NP
- NP such as LISTE
Association rules

- Based on the occurrence of terms in the same document
  \[ \text{Term}_1, \text{Term}_2 \Rightarrow \text{Term}_A, \text{Term}_B, \ldots \]
- If terms \( \text{Term}_1, \text{Term}_2 \) appear in a document then the terms \( \text{Term}_A, \text{Term}_B \) also appear in the document
- If the terms \( \text{Term}e_1, \text{Term}e_2 \) and \( \text{Term}e_A, \text{Term}e_B, \ldots \) appear in the same documents, then it exists a semantic relation between these terms
- Weighting of the rules with confidence measure "fidelity" of the terms in the analysed documents
- Several types of relations
- No automatically assigned types: approach requires an interpretation of the results and a manual analysis
Hyperonym

*histamine* is-a *biogenic amine*

Sibling

*spermidine* is-a-sibling *putrescine*

Inter-hierarchical relations

*acids* is-transformed-in *esters*

*silica* used-for *chromatography*
Distributional analysis

- Based on Zellig Harris work
- Meaning of the terms (lexical units)
  \[\leftrightarrow\] contextual proximity
- Syntactic dependency relations
  predicate-argument relation object of, subject of, etc.
- Use of lexical clues
- Results:
  - Definition of term classes
  - Semantic similarity within the classes
    - large : small, important, major, great, various, main,
      different, field, new
    - patient : case, group, child, day, treatment, woman
  - but various types of relations
Manual identification of the type of the relations

- antonymy: large, small
- synonymy: large, important, great
- meronymy: patient, group
- hyperonymy: patient, child, woman
- inter-hierarchical relations: patient, treatment
Lexical inclusion

- **Objective:** identification of hierarchical relations between terms
- **Hypothesis:** lexical inclusion
  if a term is included in another term, there is a semantic relation between them
- **Inclusion:**
  - at the string level (a string is included in another)
  - use of syntactic relation (syntactic head of a term)
Lexical inclusion

- Different types of relations
  - Hyperonymy:
    
    venous oxygen saturation is-a oxygen saturation
    poumon eosinophile is-a eosophilie
    respiratory tract neoplasm is-a neoplasm

    → Exception: Generalized Lie theory / Lie theory

- Inter-hierarchical relations
  
  right ventricular failure – failure
  differential white blood cell count – white blood
Substitution rules

- Objective: acquisition of synonymy relations between multi-word terms
- Based on the compositionality principle (Partee, 1984)
- Hypothesis: compositional propagation of synonymy relations
- Two multi-word terms are assumed to be synonymous if their components are identical or synonymous
- Exploitation of existing lexical resources thesaurus, general language dictionaries
Substitution rules

Types of substitutions:

- Variation on the expansions:
  heads are identical, expansions are synonym
  nephritic syndrome / renal syndrome

- Variation on the heads:
  heads are synonym, expansions are identical
  alcohol intoxication / alcohol poisoning

- Variation on the heads and expansions:
  head and expansions are synonyms
  dibenzo-p-dioxin metabolic process / phenodioxin metabolism
Identifying synonymy relations

SynoTerm (Hamon, 2000)

Compositional propagation of synonymy relations

- Specialised or general information (elementary synonymy relations) as a bootstrap
- Use of lexical or terminological resources
- Inference of synonymy relations between multi-word terms
  Elementary relation: *catabolism / degradation*

→ *xanthophyll catabolism / xanthophyll degradation*

Additional results:

- General language resource: good recall, low precision
- Thesaurus, or very specialised resources: low recall but high precision

→ Increasing the results by combining various types of resources
Jacquemin (1999)
Term recognition and morpho-syntactic variants
- Based on a robust and shallow syntactic analyser
  Lexicalised Tree Adjoining Grammar
- Recognition of controlled terms in corpora
  Identification of term variants
- Definition of a set of meta-rules
  describing various phenomena of term variation
- Formalism: Head-Driven Specialised Grammar (HPSG)
  and unification of feature structures
Morpho-syntactic variation:

- **Syntactic variants**
  - *body and organ weights*
  - coordination term variation of *body weight*

- **Morpho-syntactic variants**
  - *perforated ulcer* nominalised variants of *ulcer perforation*
  - *intestine perforation / intestinal ulcer perforation*

- **Syntactic variants**
  - **Permutation:**
    - *Commission request / requests of the European Commission*
    - *brain development / development of the brain*
  - **Insertion:**
    - *animal study / animal toxicity studies*
    - *development disorder / developmental coordination disorder*
Faster

- Tool used for controlled and free indexing of documents
- Acquisition of a large amount of terms (even if it is not the first objective)
- Controlled indexing: Use of a list of terms as an index
- Enrichment of the list: exploitation of this term list to recognise term variants
Conclusion

Presentation of computer-based approaches to help the building of terminologies (term extraction and relation acquisition)
But, still a lot of work to improve the results:

- **Term extraction:**
  - How to rank the candidate terms:
    - statistical information and measures
    - study of the syntactic network of the terms

- **Relation acquisition:**
  - automatically identify types of the relations
  - better exploitation of the semantic context

- Combination of several approaches
- Identification of the preferred term
- ...

Computational Terminology
T Hamon
Conclusion

Presentation of computer-based approaches to help the building of terminologies (term extraction and relation acquisition)
But, still a lot of work to improve the results:

- **Term extraction:**
  - How to rank the candidate terms:
    - statistical information and measures
    - study of the syntactic network of the terms

- **Relation acquisition:**
  - automatically identify types of the relations
  - better exploitation of the semantic context

- Combination of several approaches
- Identification of the preferred term
- ...

Tack!