mig

Unité Mathématique, Informatique et Génome

Applications and methods in *text-mining* for biology

Transys training session

Jouy-en-Josas – 1st June 2009

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

INRA

Outline of the talk

- Example of Information Retrieval application - *semantic search* of bibliography in the *Microbiology* domain

- Beyond IR in Biology, strong needs for technical and scientific content analysis.

- Specific-domain knowledge acquisition, a well-known bottleneck

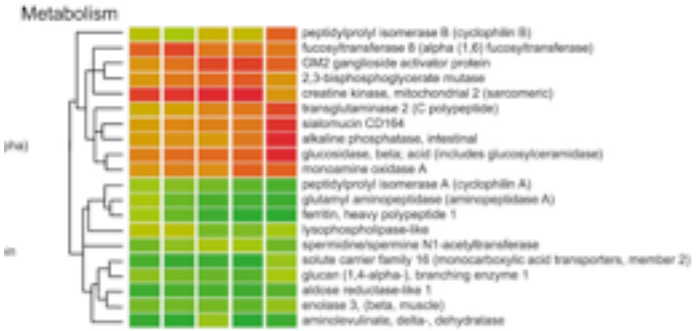
- Machine Learning from corpus-based examples, several goals, several technologies

  … at various distances to technological transfert

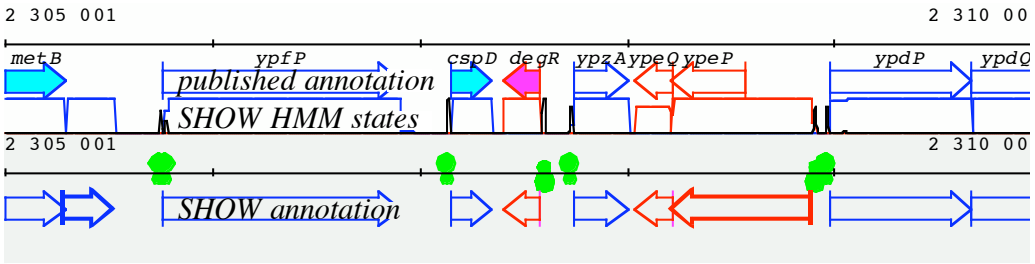
- Two examples of tasks
  - Gene / protein / species name recogition (Named entity recognition)
  - Acquisition of thesaurii

Textual documents, a major a source of information in biology

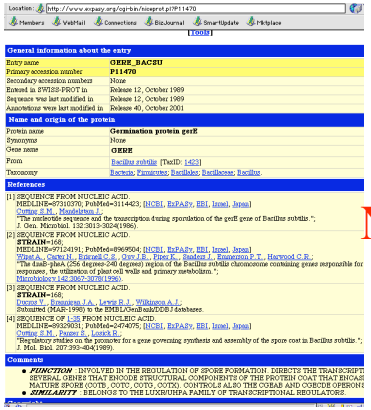
Large-scale experiments
sequencing, gene expression



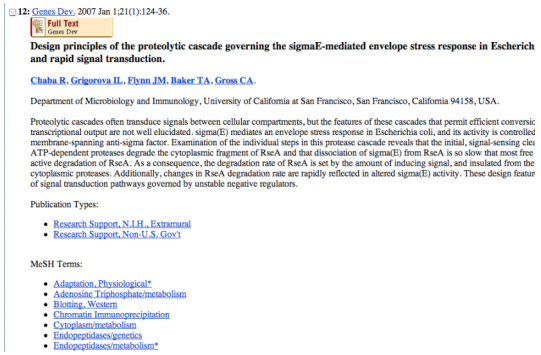
Annotation
of sequences
(genes, function)



Predictive models from
computer science and
mathematics
inférence of network regulation
motif discovery



Texts in Natural Language



Design principles of the proteolytic cascade governing the sigmaE-mediated envelope stress response in Escherich and rapid signal transduction.

Chaba R, Grigorova IL, Flynn JM, Baker TA, Gross CA.

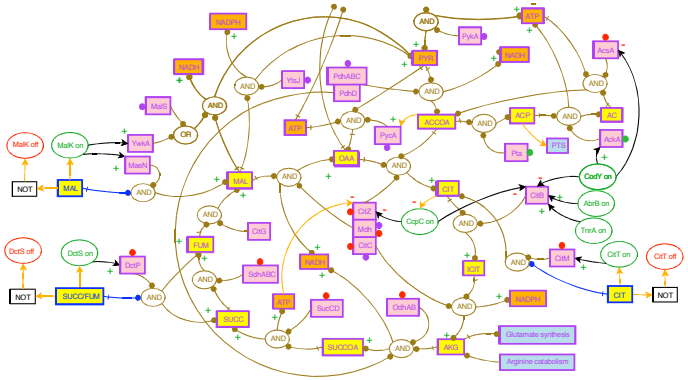
Department of Microbiology and Immunology, University of California at San Francisco, San Francisco, California 94158, USA.

Publication Types:

- Research Support, N.I.H., Extramural
- Research Support, Non-U.S. Gov't

MeSH Terms:

- Adaptation, Physiological*
- Adenosine Triphosphate/metabolism
- Blotting, Western
- Chromatin Immunoprecipitation
- Cytoplasm/metabolism
- Endopeptidases/genetics
- Endopeptidases/metabolism*



***Bibliome*, a bioinformatics field**

Bioinformatics,
extracts, analyses, manages and
proposes interpretative models from data, for the understanding of biological processes.

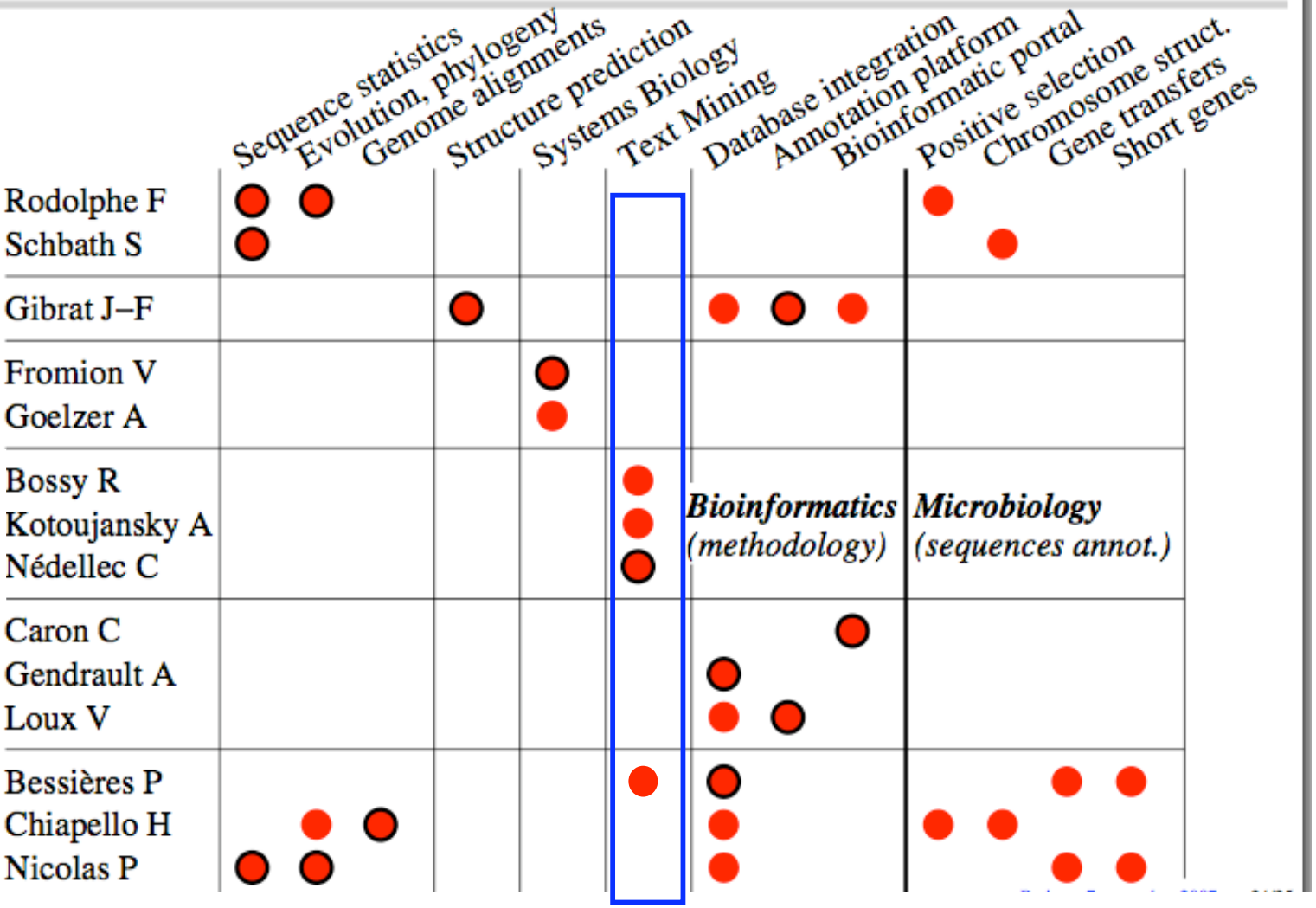
Texts,
a gold mine of information,

… for who knows
- *How* to search,
- *Where* searching,
- *What* searching

mig

Unité Mathématique, Informatique et Génome

MIG, *hub* bioinformatique et microbiologique



Aubin S., Jourde J.
Lemaçon A.
Papazian F.,
Makuntima S.

mig
Unité Mathématique, Informatique et Génome

Scientific documents on stress of *Bacillus subtilis*

stress factor Bacillus subtilis

Scholar **Tous les articles - Articles récents** Résultats **1** - **10** sur un total d'env

Heat-shock and general **stress** response in **Bacillus subtilis** M Hecker, W Schumann, U Volker - Molecular Microbiology, 1996 - Blackwell Synergy **...** of the adaptional network of a non-growing cell of **Bacillus subtilis**. **... stress** proteins that may confer specific protection against a particular **stress factor**. **...** Cité 388 fois - Autres articles - Recherche sur le Web - Les 6 versions

… the sigB operon of **Bacillus subtilis** that control activity of the general **stress factor** sigma B in … AA Wise, CW Price - Journal of Bacteriology, 1995 - Am Soc Microbiol ... positive bacterium **Bacillus subtilis** is an alternative transcription **factor** activated by a variety of environm té 103 fois - Autres articles - Recherche sur le Web - Les 5 versic

Looks for exact words

… specific, general and multiple **stress** resistance of growth-restricted **Bacillus subtilis** cells by the … M Hecker, U Volker - Molecular Microbiology, 1998 - Blackwell Synergy **...** A., Garsin, DA, Duncan, L., Losick, R. (1996) Role of adenosine nucleotides in the regulation of a **stress** response transcription **factor** in **Bacillus subtilis**. **...** Cité 161 fois - Autres articles - Recherche sur le Web - Les 6 versions

Stress-induced activation of the sigma B transcription **factor** of **Bacillus subtilis**. SA Boylan, AR Redfield, MS Brody, CW Price - Journal of Bacteriology, 1993 - Am Soc Microbiol **... stress** regulon controlled by us (19), and the sporulation process of **Bacillus subtilis**, controlled by a cascade of at least six different sigma **factors** (13, 36 **...** Cité 136 fois - Autres articles - Recherche sur le Web - Les 6 versions

Example: bibliographic search by information retrieval

Query: *stress Bacillus subtilis*

Full text search by: ***Google***

317 answers: *stress*, *Bacillus* et *subtilis* are viewed as simple words independently searched. Relevant papers are missed and irrelevant papers retrieved

Full text search and keywords in a hierarchical thesaurus : ***PubMed***

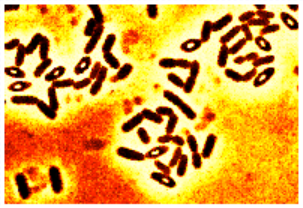
211 answers: *stress* is viewed as a simple word and *Bacillus subtilis* denotes a species according to *MeSH* thesaurus.

Full semantic search : ***BioAlvis*** *(http://genome.jouy.inra.fr/alvis/front)*

471 answers: *stress is a environment condition (e.g. oxidative stress, elevated temperature)* and *Bacillus subtilis* as a *Bacteria*.

Different results due to document collections and semantic analysis capabilities

mig
Unité Mathématique, Informatique et Génome



stress factor bacillus subtilis Search

Concept all [-]

environment condition
- stress factor
  - effect of temperature
    - elevated temperatu
      - heat-shock

Species all [+]

Bacillus subtilis

Genes all [+]

YtkD

Authors all [+]

Hecker M

Dates all [+]

2006

Journals all [+]

J Bacteriol

Query details: bacillus(lemma) subtilis(lemma) stress factor(Subtilis Molecular Biology Concept/environment conditio

21-30 among 471 results in 713 categories

Cloning in Bacillus subtilis of an extremely thermostable alpha amylase: comparison with

Cloning in **Bacillus** **subtilis** of an extremely thermostable alpha amylase: comparison with other cloned heats **subtilis**. ... licheniformis FDO2 and very similar to the restriction map of a high temperature amylase from **Ba**

Subtilis Molecular Biology Concept/species/cellular organisms/Eukaryota/Fungi-Metazoa group/Metazoa/Eumetazoa/Bilateria/Coelomata/Protostomia/Panarthropoda/Arthropoda/Mandibulata/Pancrustacea/H subtilis Subtilis Molecular Biology Concept/environment condition/stress factor/effect of temperature/elevated temper group/Metazoa/Eumetazoa/Bilateria/Coelomata/Protostomia/Panarthropoda/Arthropoda/Mandibulata/Pancrustacea/H licheniformis Subtilis Molecular Biology Concept/protein concept/protein/protein of known function/enzyme/alpha-amy function/enzyme gene/amylase gene/alpha-amylase gene restriction endonuclease Biochem Biophys Res Commun Bi Enzymes alpha-Amylase 1984 Piggott R P Rossiter A Ortlepp S A Pembroke J T Ollington J F Bacillus subtilis Bacillus li

Analysis of Bacillus subtilis tagAB and tagDEF expression during phosphate starvation ide

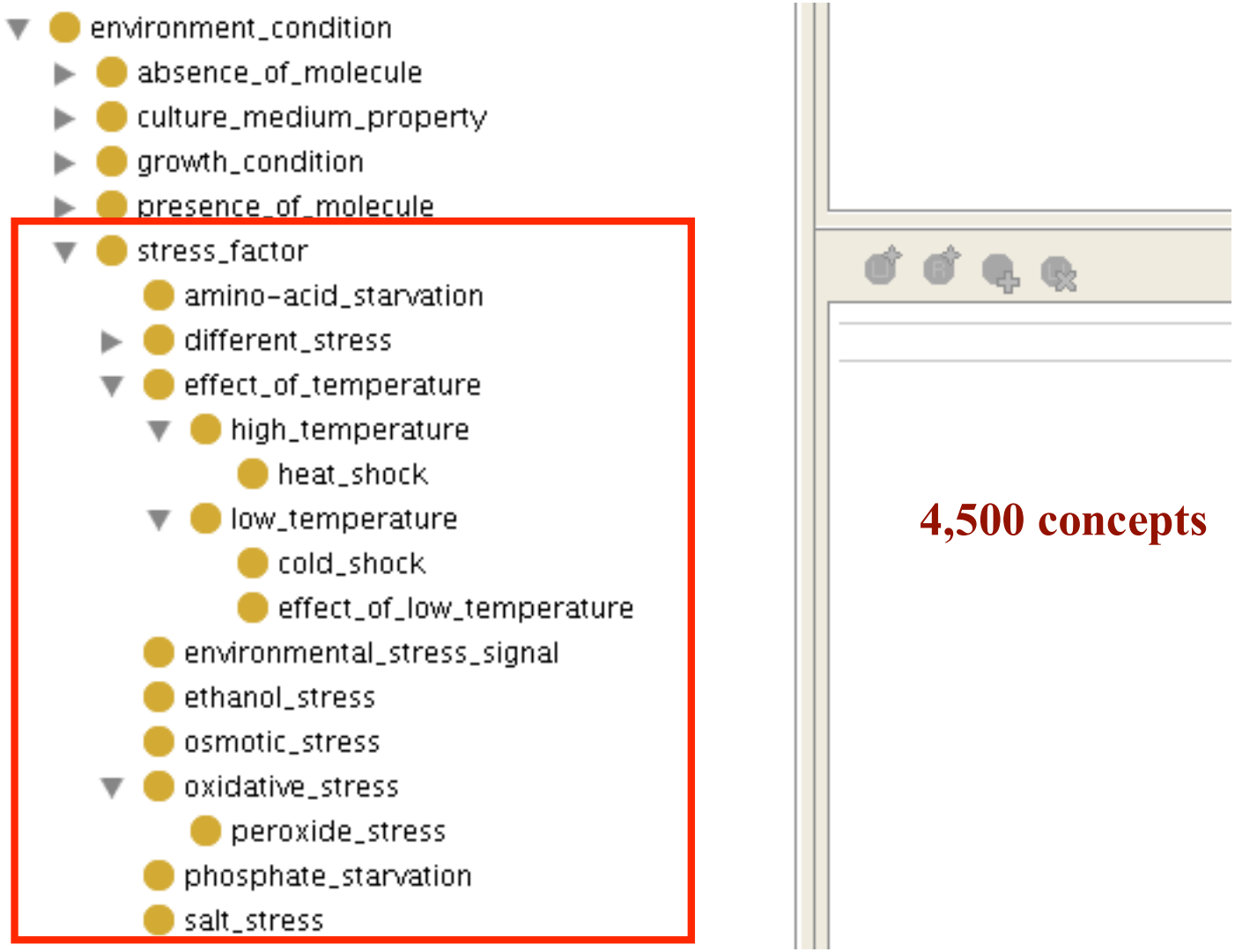
Analysis of **Bacillus** **subtilis** tagAB and tagDEF expression during phosphate starvation identifies a repressor presented here suggest that PhoP and PhoR are required for direct rep

tagAB tagDEF PhoP-P PhoP PhoR Pho Subtilis Molecular Biology Concept/gene concept/gene organization/gene Subtili region/operon Subtilis Molecular Biology Concept/environment condition/stress factor/phosphate-starvation Subtilis M condition tagAB tagDEF J Bacteriol Journal of bacteriology Bacterial Proteins Gene Expression Regulation, Bacterial Mo Eder S Hulett F M Bacillus Bacillus subtilis

A relA(S) suppressor mutant allele of Bacillus subtilis which maps to relA and responds o

A relA(S) suppressor mutant allele of **Bacillus** **subtilis** which maps to relA and responds only to carbon limita relA mutant of **Bacillus** **subtilis**. ... **subtilis** chromosome.

Extract of BioAlvis ontology on stress



mig

Unité Mathématique, Informatique et Génome

gel shift DinR binding bacillus subtilis Search

~~all~~ [+] Query details: binding(lemma) bacillus(lemma) subtilis(lemma) gel shift(term) DinR(text)

1-2 among 2 results in 87 categories

The Bacillus subtilis DinR binding site: redefinition of the consensus sequence.

Electrophoretic mobility **shift** assays revealed that highly purified **DinR** does bind to such sites locat
others for maintaining efficient **DinR binding**. ... **subtilis dinR** and recA **DinR** boxes revealed that t

DinR recA Cheo lexA uvrC dinB tagC N4 DinR Subtilis Molecular Biology Concept/species/cellular organisms/
group/Metazoa/Eumetazoa/Bilateria/Coelomata/Protostomia/Panarthropoda/Arthropoda/Mandibulata/Pancr
subtilis Subtilis Molecular Biology Concept/protein concept/protein binding site Subtilis Molecular Biology Con
all [+] method/sequence similarity/consensus sequence gel shift J Bacteriol Journal of bacteriology Amino Acid Sec
(Genetics) 1998 Winterling K W Levine A S Yasbin R E Woodgate R Chafin D Hayes J J Sun J Bacillus subtilis

all [+] **Characterization of DinR, the Bacillus subtilis SOS repressor.**

By using electrophoretic mobility **shift** assays, we demonstrated that **DinR** interacts with the previou
box. ... **subtilis**.

all [+] SOS LexA Gly RecA DinR recA Subtilis Molecular Biology Concept/regulator/regulon Gram-negative Subtilis M
organisms/Bacteria/Proteobacteria/Gammaproteobacteria/Enterobacteriales/Enterobacteriaceae/Escherichia
Biology Concept/gene concept/genetic variation/DNA damage SOS-induction Subtilis Molecular Biology Conc
all [+] Bacillus subtilis Bacterial Proteins Consensus Sequence Molecular Sequence Data 1997 Winterling K W Levin

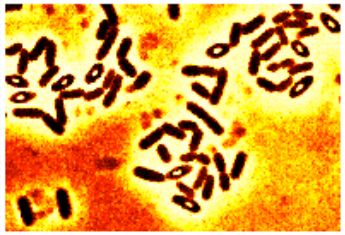
Extract of *BioAlvis* terminology

5	transcription repression	repression of the transcriptic	repression of gene transcri	repression of transcription	
6	restriction mapping	restriction endonuclease site	restriction enzyme mappin	restriction enzyme site mapping	restriction site m
7	rho-independent termir	rho-independent transcriptio	rho-independent transcript	rho-independent terminator of transcription	
8	S1 mapping	S1 transcript mapping	S1 protection mapping	S1 nuclease protection mapping	S1 nuclease tran
9	gene sequence	sequence of a gene	sequence of the gene	sequence deduced from gene	sequence of gen
0	promoter sequence	promoter DNA sequence	sequence for promoter	sequence of the promoter	sequence for the
1	Southern analysis	Southern blot hybridization a	Southern hybridization ana	Southern blotting analysis	Southern blot an
2	polar septum	polar division septum	spore septum	sporulation septum	
3	high similarity	high degree of similarity	high level of similarity	strong similarity	
5	heat-shock	temperature upshift	thermal upshock	temperature upshock	
7	terminator structure	structure of the terminator	terminator RNA structure	termination structure	
9	transcription start site	start site of transcription	site of transcription initiati	transcription initiation site	
0	gene transcription	transcription from gene	transcription of the gene	transcription of gene	transcription of t
1	transcription factor	factor required for efficient t	factor in the transcription	transcriptional factor	
2	transcription inhibition	inhibition of the transcriptior	inhibition of transcription	transcriptional inhibition	
3	transcription-initiation	initiation of RNA transcriptio	initiation of transcription	transcriptional initiation	
4	transcription regulation	regulation of gene transcript	regulation of transcription	transcriptional regulation	transcriptional re
5	transcription start	transcription initiation start	start of transcription	transcriptional start	
6	transcription terminatio	termination of DNA transcrip	termination of transcriptior	transcriptional termination	
7	Western analysis	Western immunoblot analysi	Western blot analysis	Western immunoblot assay	Western immunc

15,000 terms

mig

Unité Mathématique, Informatique et Génome



spolIIG sporulation bacillus subtilis Search

1,689,244 genes / proteins

748,262 species

Concept ~~all~~ [+]

- gene function
  - sporulation gene
    - forespore gene
      - forespore-spec
- gene concept

Species all [+]

Bacillus subtilis

Genes all [+]

sigG

Authors all [+]

Setlow P

Dates all [+]

2005

Query details: bacillus(lemma) subtilis(lemma) sporulation(Subtilist functional classification/cell envelope and

1-10 among 65 results in 611 categories

Expression of the Bacillus subtilis spoIVB gene is under dual sigma F/sigma G cont

However, during **sporulation**, only sigma G directs significant levels of spoIVB expression.

sigG sigF spoIVB Subtilist functional classification/information pathway/RNA synthesis/initiation of RNA synth organisms/Eukaryota/Fungi-Metazoa group/Metazoa/Eumetazoa/Bilateria/Coelomata/Protostomia/Panarthropoda/Arthropoda/Mandibulata/Pancru subtilis Subtilist functional classification/information pathway/RNA synthesis/initiation of RNA synthesis/sigF S function/enzyme/polymerase/RNA-polymerase Subtilist functional classification/cell envelope and cellular pro Factor Gene Expression Regulation, Bacterial Transcription Factors Base Sequence Molecular Sequence Data

Analysis of the interaction between the transcription factor sigmaG and the anti-s

The activation of sigma(G), a transcription factor, in **Bacillus subtilis** is coupled to the completion o

SpoIIAB SpoIIAA sigG sigF Subtilist functional classification/cell envelope and cellular process/sporulation/Sp Subtilis Molecular Biology Concept/regulator/transcription factor/anti-sigma-factor Subtilis Molecular Biology process/sporulation/SpoIIAA Subtilis Molecular Biology Concept/cell concept/cell-cycle/sporulation J Bacteriol Bacterial 2003 Evans Louise Errington Jeff Feucht Andrea Clarkson Joanna Yudkin Michael D Bacillus subtilis

Transcription of spoIVB is the only role of sigma G that is essential for pro-sigma K

Activation of pro-sigma K processing in the mother cell at late stages of **sporulation** in **Bacillus sub**

Full-text semantic search in *BioAlvis*

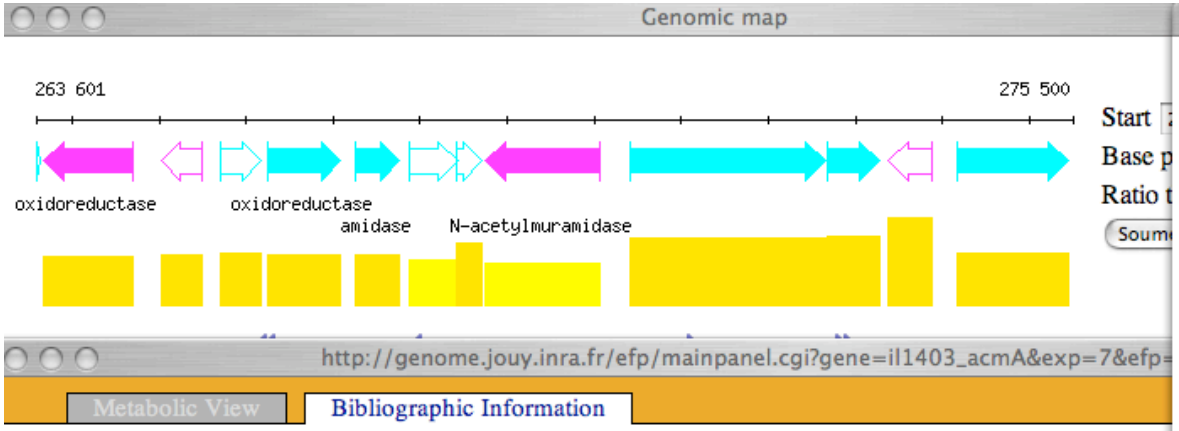
Semantic search interprets user queries by using

- Synonym terms (*spoIIIG / sigma(G)* *gel electrophoresis shift / gel shift*)
- More specific terms (*stress factor* → *phosphate starvation*)
- Relations (*stress factor* applied *to Bacillus subtilis*)

***Full-text* search of precise information** with **high recall**

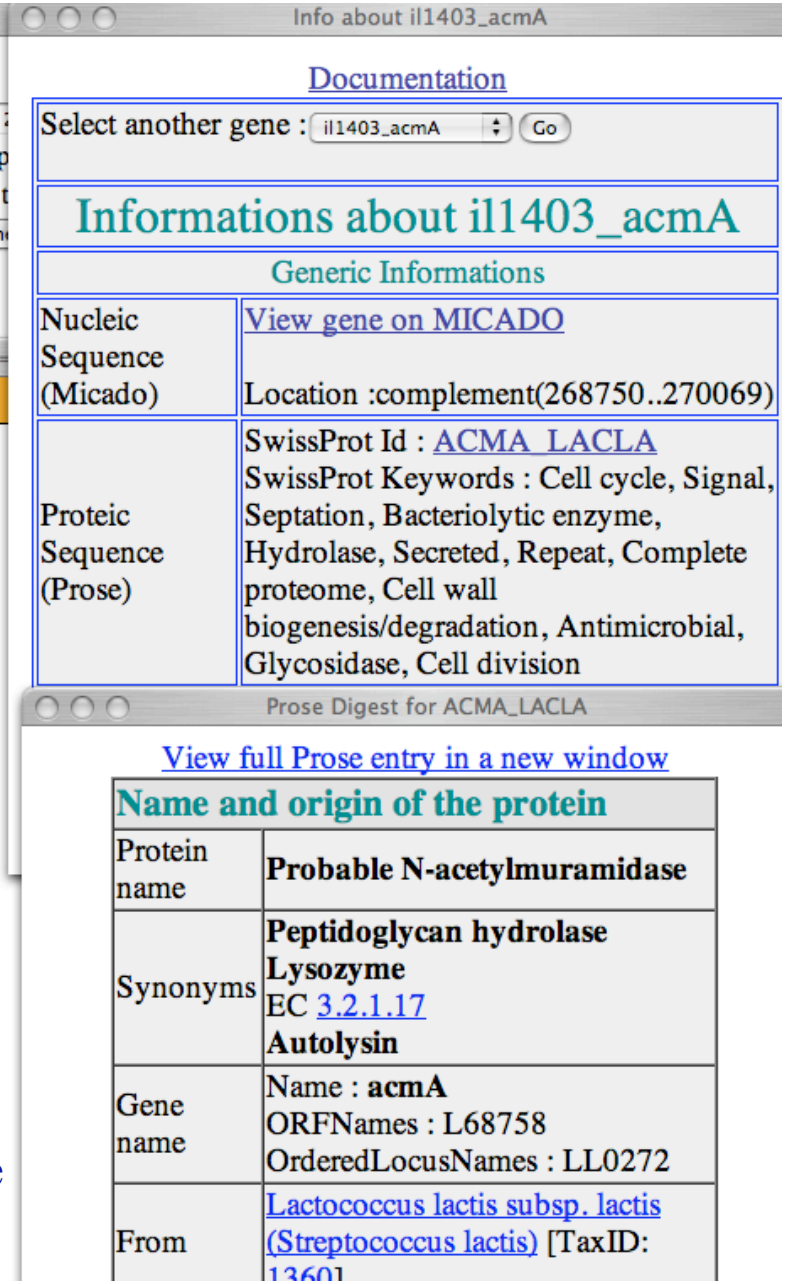
- Relies on **semantic analysis** of the content
- Requires **specific resources**, *terminology, thesaurii, variation & disambiguation rules*
- Not available in the general case and **costly to acquire**.

Bibliome research group of MIG lab. (INRA) develops **methods for knowledge acquisition** from textual documents in biology and integrate them in *on-line services*.



Sentence	PMID
Of the cell surface display constructs with the **AcmA** anchor, only those with the longest PrtP spacer regions resulted in efficient **binding** of recombinant L. lactis cells to porcine intestinal **epithelial cells**	15066797 Abstract Annotation
Double **acmA** ponA mutants displayed **increased** **adhesion** and biofilm-forming capacity	12366846 Abstract Annotation
Double **acmA** ponA mutants displayed **increased** adhesion and **biofilm-forming** capacity	12366846 Abstract Annotation

Lists : Home | List of agents | List of targets | List of actors



Gene interactions in *Express-Fingerprints* integrated interface

Gene interactions in *Express-Fingerprints* integrated interface

Complementary types of information on a given *Lactococcus lactis* gene called *acmA*

1. gene annotations and expression levels (genomic map frame),
2. biochemical function (*hydrolase of the cell wall*) from SwissProt (down right frame),
3. biological function (*adhesion to ephitelial cell* and *biofilm-forming*) from PubMed abstracts

Contribution of mining text from publications, here correlating the biochemical and the biological functions.

Gene interaction, an example of knowledge chunk to be extracted

Scientific paper on *Bs* regulation [Zheng et al., 92], among thousands

In addition,
GerE stimulates cotD transcription
[..]

?

Automatically building network regulation?

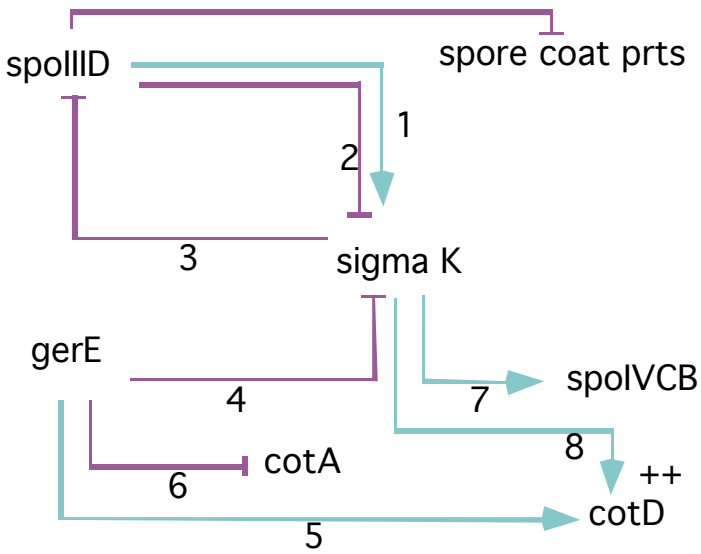
[*Manine*, Alphonse, Bessières, Nedellec]

positive interact.

negative interact.

1. SpoIIID is needed to produce sigma K

2. SpoIIID is capable of altering the specificity of RNAP-sigma K



3. Production of sigma K leads to a decrease in the level of spoIIID

4. GerE profoundly inhibits in vitro transcription of sigK encoding sigma K

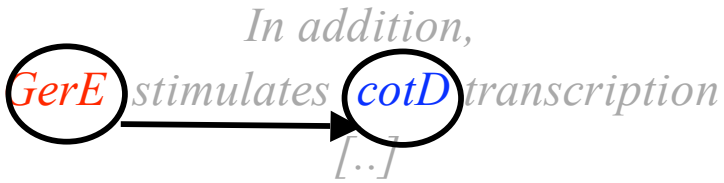
5. GerE stimulates cotD transcription

6. ... and inhibits cotA transcription.

7. sigma K has been found that causes weak transcription of spoIVCB

8. ... and strong transcription of cotD.

From text analysis



What should the computer know?

That proteins activate gene transcription
That genes *express* proteins
The names of genes and proteins
Formulation of *transcription activation* and gene *expression*
...

→ Critical problem of text analysis, knowledge acquisition and modeling

Example of extraction rule

GerE stimulates cotD *transcription […]*

The rule applies to (*GerE*, *cotD*)

Positive_interaction (X, Z):-

is-a(X, protein), subject(X, Y), is-a(Y, pos_interaction), Obj(T, Y), is-a(T, transcription), is-a(Z, gene), comp:N-N(T,Z)

Interpretation

If the subject X of a *positive interaction* verb Y is a *protein name*, the object T is a *transcription*, the complement of which, Z, is a *gene name*.

Then, X is the agent and Z is the target of the genic interaction.

Even for simple cases, syntactic and semantic analysis is needed

Semantic analysis

- Relevant for many document-based aplications

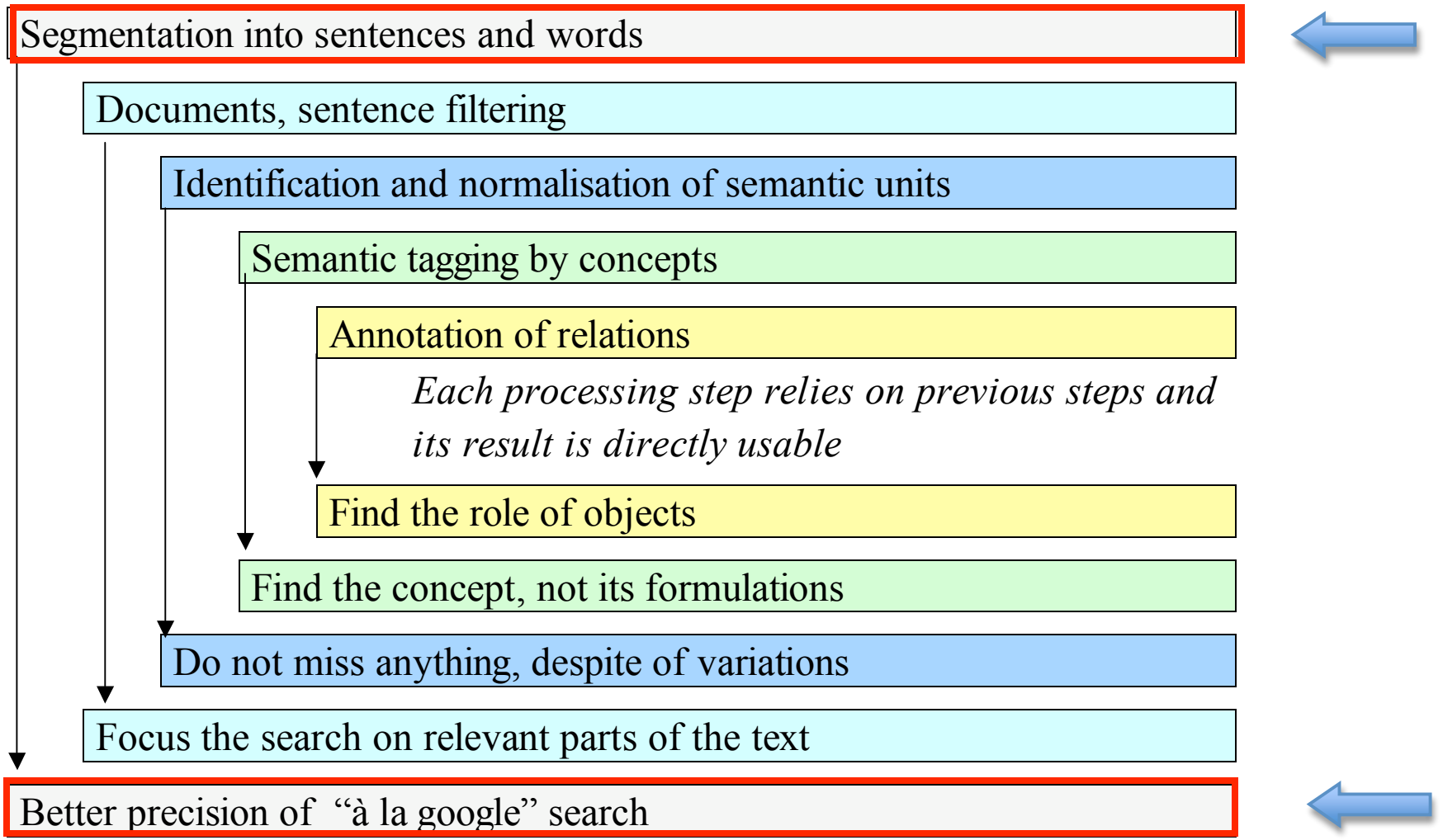
  (*e.g.* Information Extraction, Information Retrieval, Question/Answering, Summarization)

- Several intermediate computational linguistic analysis steps are needed

- Each step requires specific domain knowledge

- Text analysis methods and domain knowledge are highly reusable in various applications

Semantic analysis steps and their role in end-user applications



Segmentation into words and sentences, for a better precision

– We have previously reported that YaaH and YrbA are spore proteins of Bacillus subtilis that are required for spore resistance and/or germination and that they have a motif conserved among so-called cell wall binding proteins [Kodama et al. (1999) J. Bacteriol. 181, 4584-4591, Takamatsu et al. (1999) J. Bacteriol. 181, 4986-4994].

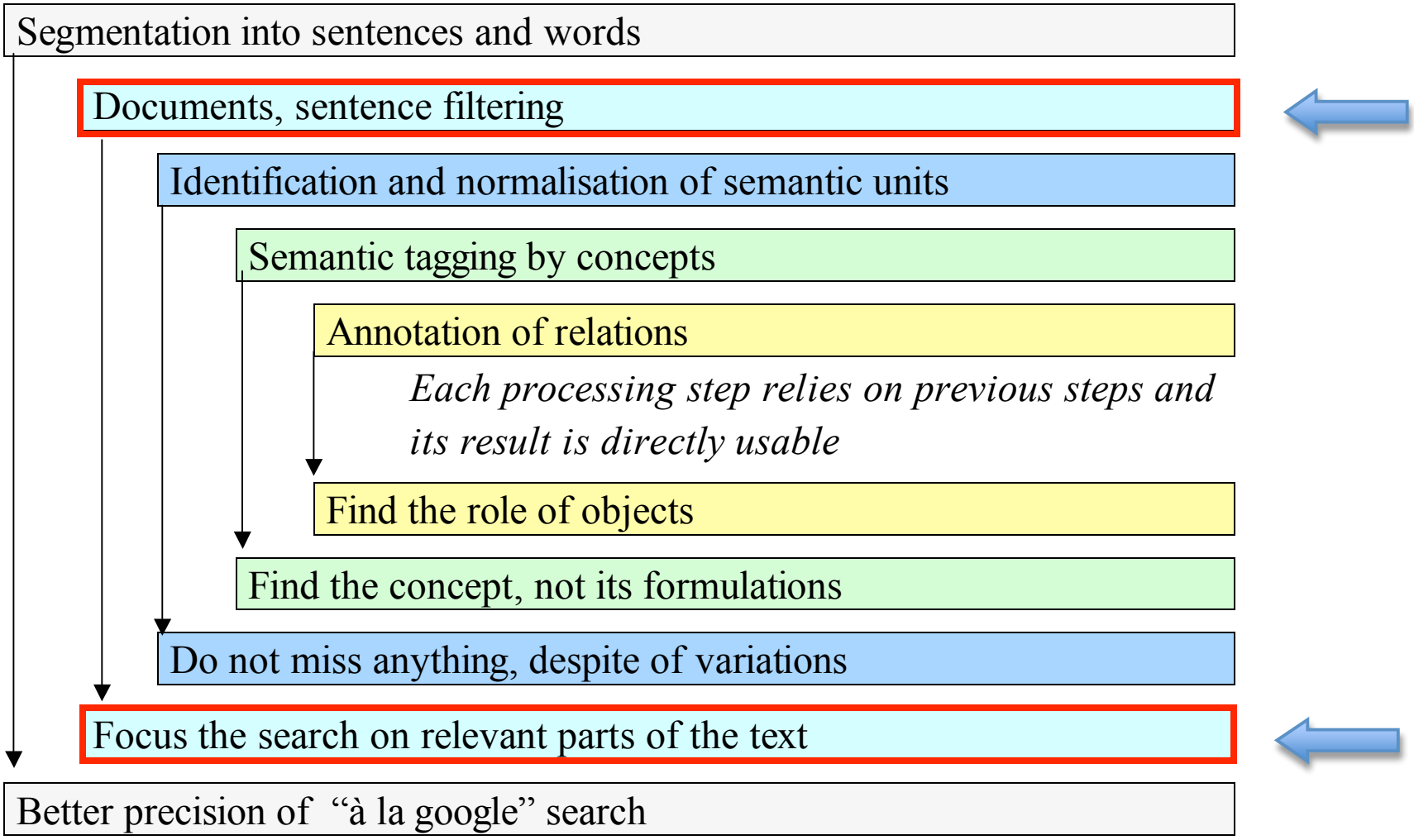
*Learn how to segment sentences on **dots**?*

identified a region of the C-terminal half of Spo0A that is highly conserved among species of endospore-forming Bacillus and Clostridium and which encodes a putative helix-turn-helix DNA-binding domain.

*Learn how to segment words on **hyphens**?*

Each domain has its own needs of specific knowledge, patterns and lexicon

Semantic analysis steps and their role in end-user applications



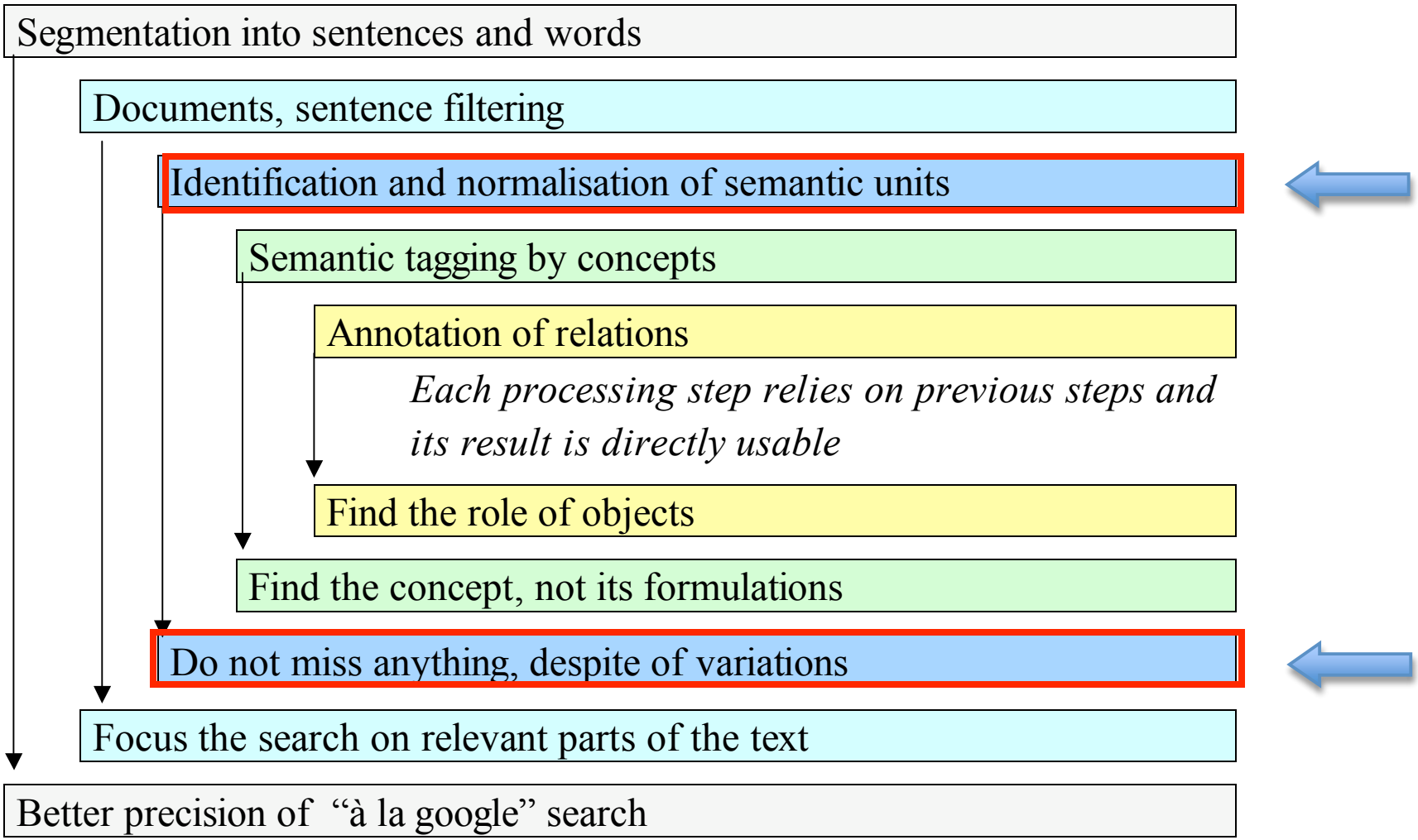
Document and sentence filtering → focus the search

STFILTER [Abdel Vetah et al., 2001]

UI - 99175219
AU - Ichikawa H
AU - Halberg R
AU - Kroos L
TI - Negative regulation by the Bacillus subtilis GerE protein.
..
PT - JOURNAL ARTICLE
TA - J Biol Chem
AB - GerE is a transcription factor produced in the mother cell compartment of sporulating Bacillus subtilis. It is a critical regulator of cot genes encoding proteins that form the spore coat late in development. Most cot genes, and the gerE gene, are transcribed by sigmaK RNA polymerase. Previously, it was shown that the GerE protein inhibits transcription in vitro of the sigK gene encoding sigmaK. Here, we show that GerE binds near the sigK transcriptional start site, to act as a repressor. A sigK-lacZ fusion containing the GerE-binding site in the promoter region was expressed at a 2-fold lower level during sporulation of wild-type cells than gerE mutant cells. Likewise, the level of SigK protein (i. e. pro-sigmaK and sigmaK) was lower in sporulating wild-type cells than in a gerE mutant. These results demonstrate that sigmaK-dependent transcription of gerE initiates a negative feedback loop in which GerE acts as a repressor to limit production of sigmaK. In addition, GerE directly represses transcription of particular cot genes. We show that GerE binds to two sites that span the -35 transcription. The upstream GerE-binding site was required for activation but not for repression. These results suggest that a rising level of GerE in sporulating cells may first activate cotD transcription from the upstream site then repress transcription as the downstream site becomes occupied. Negative regulation by GerE, in addition to its positive effects on transcription, presumably ensures that sigmaK and spore coat proteins are synthesized at optimal levels to produce a germination-competent spore.

AD - Department of Biochemistry, Michigan State University, East Lansing, Michigan 48824, USA.PMID- 0010075739
EDAT- 1999/03/13 03:11
SO - J Biol Chem 1999 Mar 19;274(12):8322-7

Semantic analysis steps and their role in end-user applications



Identification and normalisation of *semantic units* → Nothing missing

CoCitation [Bessières-Aubin, 07], **RenBio** [Bossy *et al.*, 06], **Yatea** [Hamon-Aubin, 06], **Fastr** [Jacquemin]

/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/ ./

Named entities

Bacillus subtilis = *Bs* = *B. subtilis* = *B.subtilis* — Type: Species

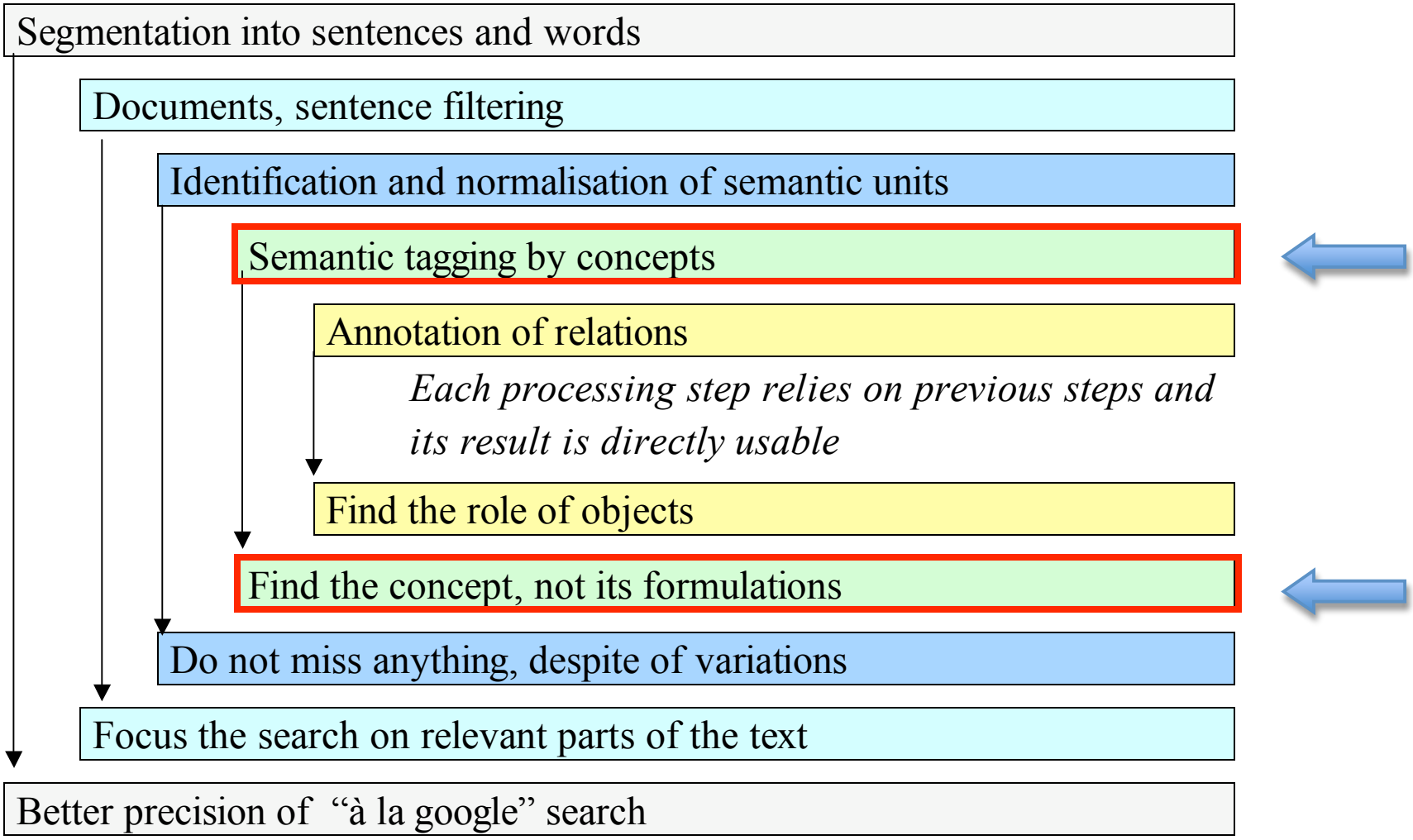
sigma(K) = *sigma (K)* = *sigmaK* = *sigma K* — Type: Transcription factor

Terms

RNA polymerase = *ribonucleic acid polymerase*

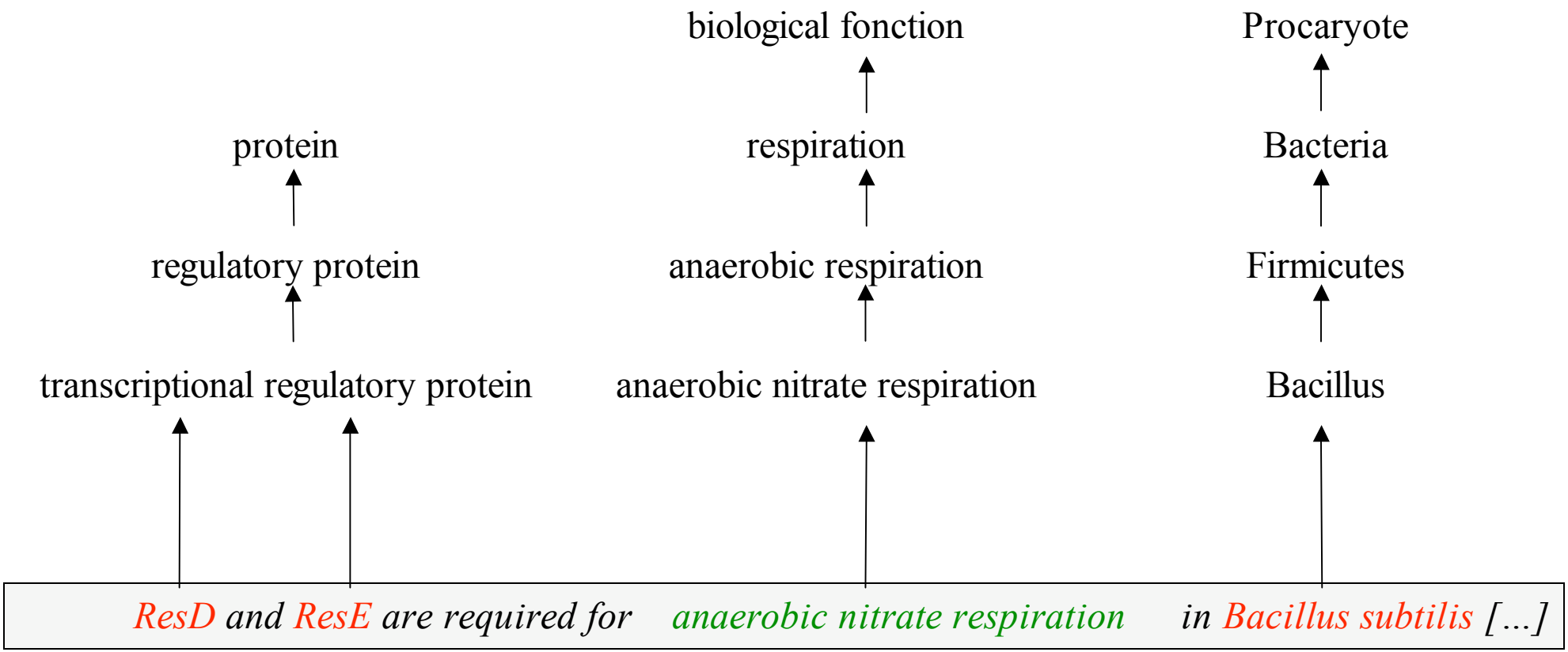
transcription factor = *factor of transcription*

Semantic analysis steps and their role in end-user applications

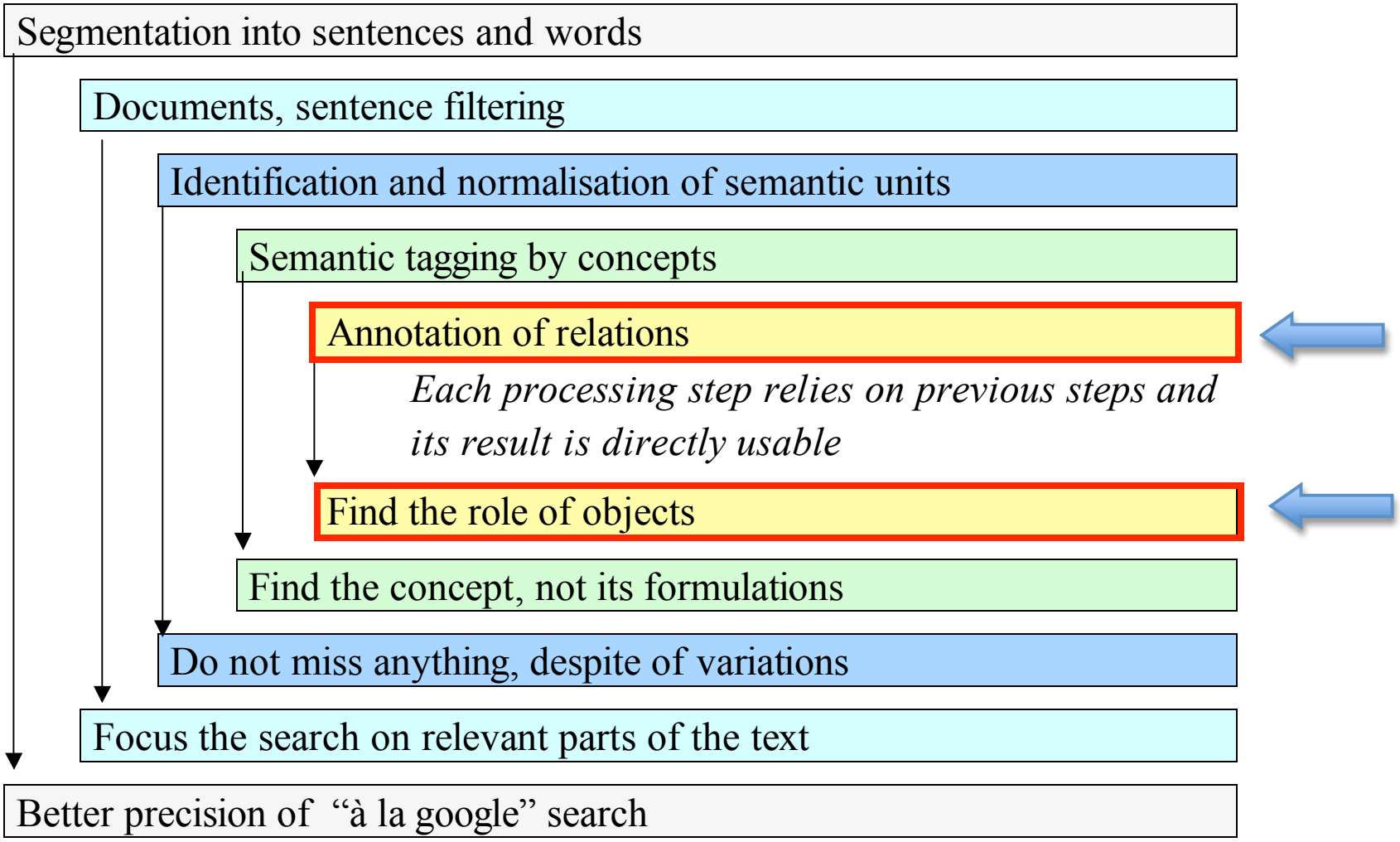


Semantic tagging → Find the concept independently of its formulation

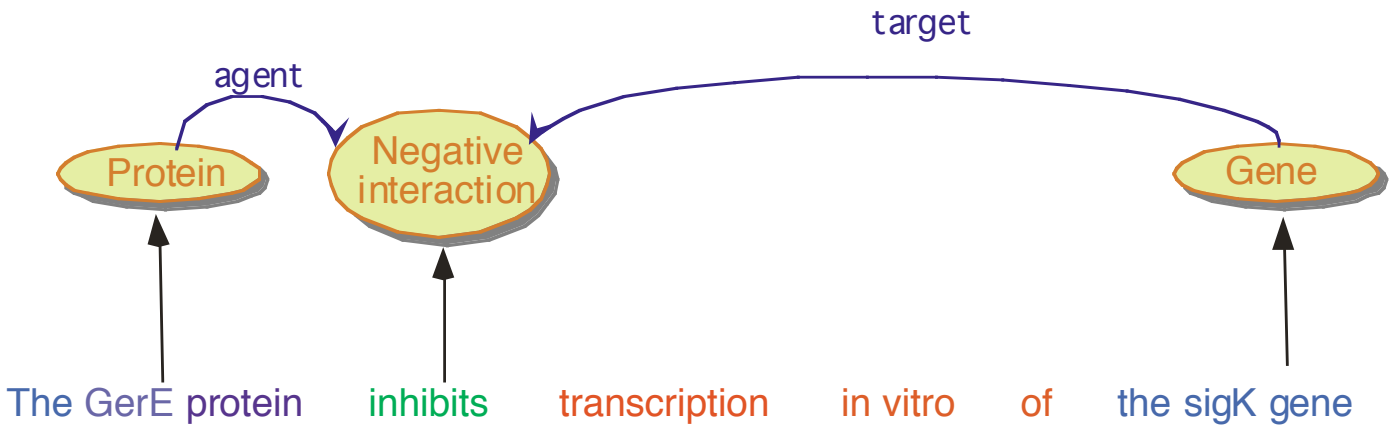
SemTag [Bossy, 06], **Asium** [Nedellec & Faure, 98], **LP2LP** [Aubin & Alphonse., 05],



Semantic analysis steps and their role in end-user applications

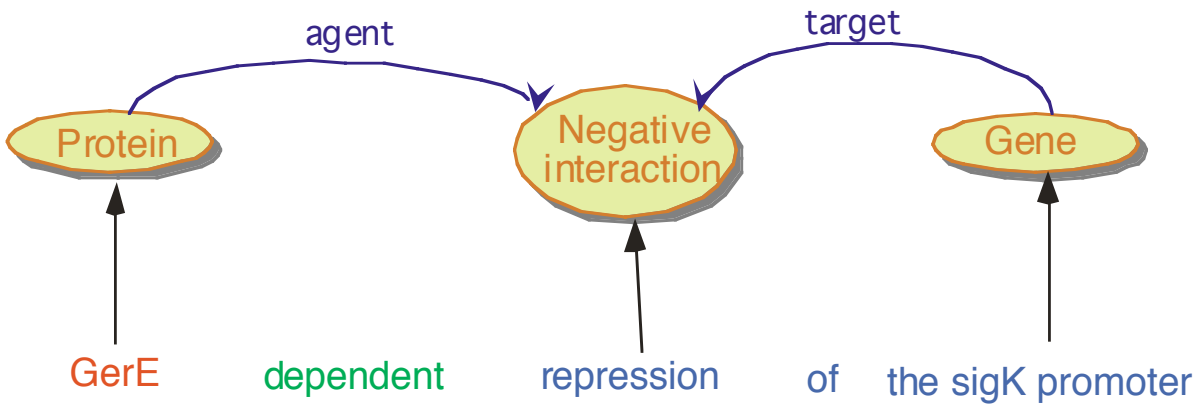


Annotation of relations → Find the role of objects



CadInteract
[Bessières et al, 06]

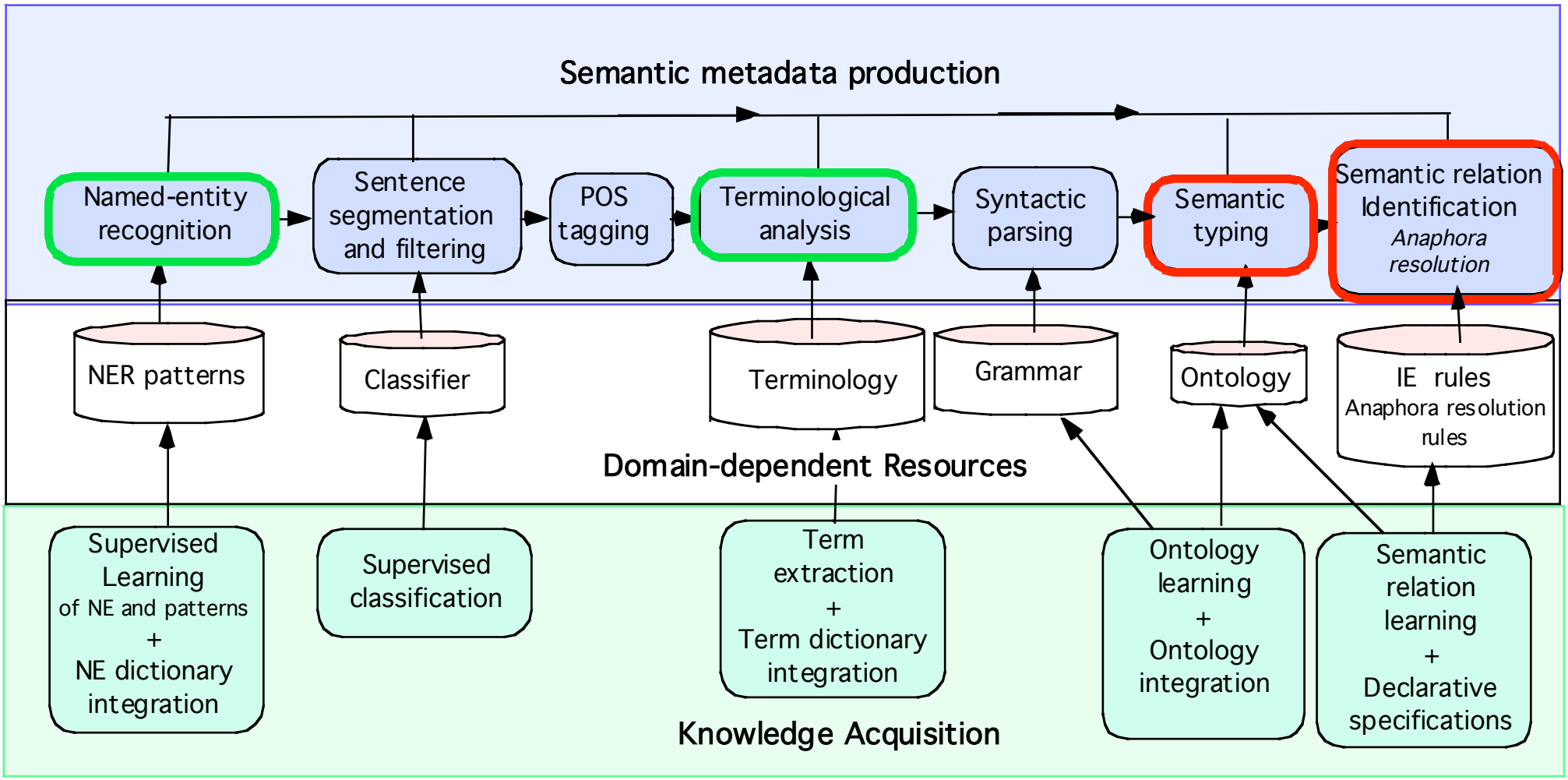
LP-Propal
[Manine et al, 05]



A same interpretation for different formulations.

Methods

BioAlvis, a flexible architecture for *text-mining* in Biology



(1) Operational services *vs* (2) research

1. *BioAlvis* and *CoCitation* on-line services available for information search in molecular bacteriology

Focused literature mining, compared to Google, PubMed or WoS.

2. More accurate results with specific knowledge bases

- Improving indexing of ***gene and protein names***, by handling
  - Gene renaming *(sigB rpoF)*
  - Typographic variations of names *(sigma B sigma(B) sigmaB*
  - Word sense ambiguity (***ara***? *bird*, *promoter*, *operon? Or Japanese biologist*)
  - Attachment of the name to the species (*GuaA* of *Bs* or *Ec* ?)
- Improving query expansion about *specific* issues (e.g. *stress*) by enriching the ***terminology and the ontology***

Operational services *vs* research

3. Automating the extraction of relation from text

- Generality relation
  - *phosphate starvation* is-a type of *stress*

- Gene interaction,
  - *GerE stimulates cotD transcription*

- Interaction of the bacteria with its environment
  - *M. tuberculosis invades and survives within human type II alveolar epithelial cells*

- Sequence homology
  - *two Bacillus subtilis genes, yklA and ykzA, which are homologous to the partially RpoS-controlled osmC gene from Escherichia coli)*

Gene and protein names recognition

MIG Mathématique Informatique & Génome

INRA

Cocitations for gene or protein BdbC with BdbD

Home page Back

Bacillus subtilis

Sentence	PMID
Four enzymes of this type, termed **BdbA**, **BdbB**, **BdbC**, and **BdbD**, have been identified in the Gram-positive eubacterium Bacillus subtilis.	11872755
BdbC and **BdbD** have been shown to be critical for the folding of a protein required for DNA uptake during natural competence.	11872755
BdbC and **BdbD** are orthologs of enzymes known to be involved in extracytoplasmic disulfide bond formation.	11744713
Taken together, these observations imply that in the absence of either **BdbC** or **BdbD**, **ComGC** is unstable and that **BdbC** and **BdbD** catalyze the formation of disulfide bonds that are essential for the DNA binding and uptake machinery.	11744713
Consistent with this, **BdbC** and **BdbD** are needed for the secretion of the Escherichia coli disulfide bond-containing alkaline phosphatase, **PhoA**, by B. subtilis.	11744713
BdbC and **BdbD** are thiol-disulfide oxidoreductases.	11844773
Mutations in the thiol-disulfide oxidoreductases **BdbC** and **BdbD** can suppress cytochrome c deficiency of **CcdA**-defective Bacillus subtilis cells.	11844773

Gene and protein names recognition

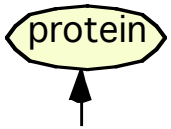
Automatic recognition from **dictionaries** and **rules** learned from training corpora (C4.5).

⇨ 93% for protein and gene names (including new names, variation, synonymy and disambiguation) by *RenBio* [Nedellec et al., 06].

⇨ 75% without corpus-based machine learning

An example of learned rule (=naming convention)

A *name*, followed by *protein* word, 4 letters long, starting and ending with an upper case letter, is a protein name.



The GerE protein inhibits transcription ...

Dictionary extract

sigma G	*sigma(G)*	*sigmaG*	*SpoIIIG*	*SpolllG*
oxaA1	*spoIIIJ*	*spo0J87*		
rpsZ	*rpsN*	*rpsN1*	*rpsNA*	

Acquisition strategy for gene/protein name recognition

Goal: improving recognition of new names and ambiguous names

Six steps

1. Definition of the needs and the acquisition task
2. Selection of a relevant training corpus
3. Automatic annotation by the current resources
4. Manual correction of the annotation
5. New recognition rule learning
6. Integration in the target service

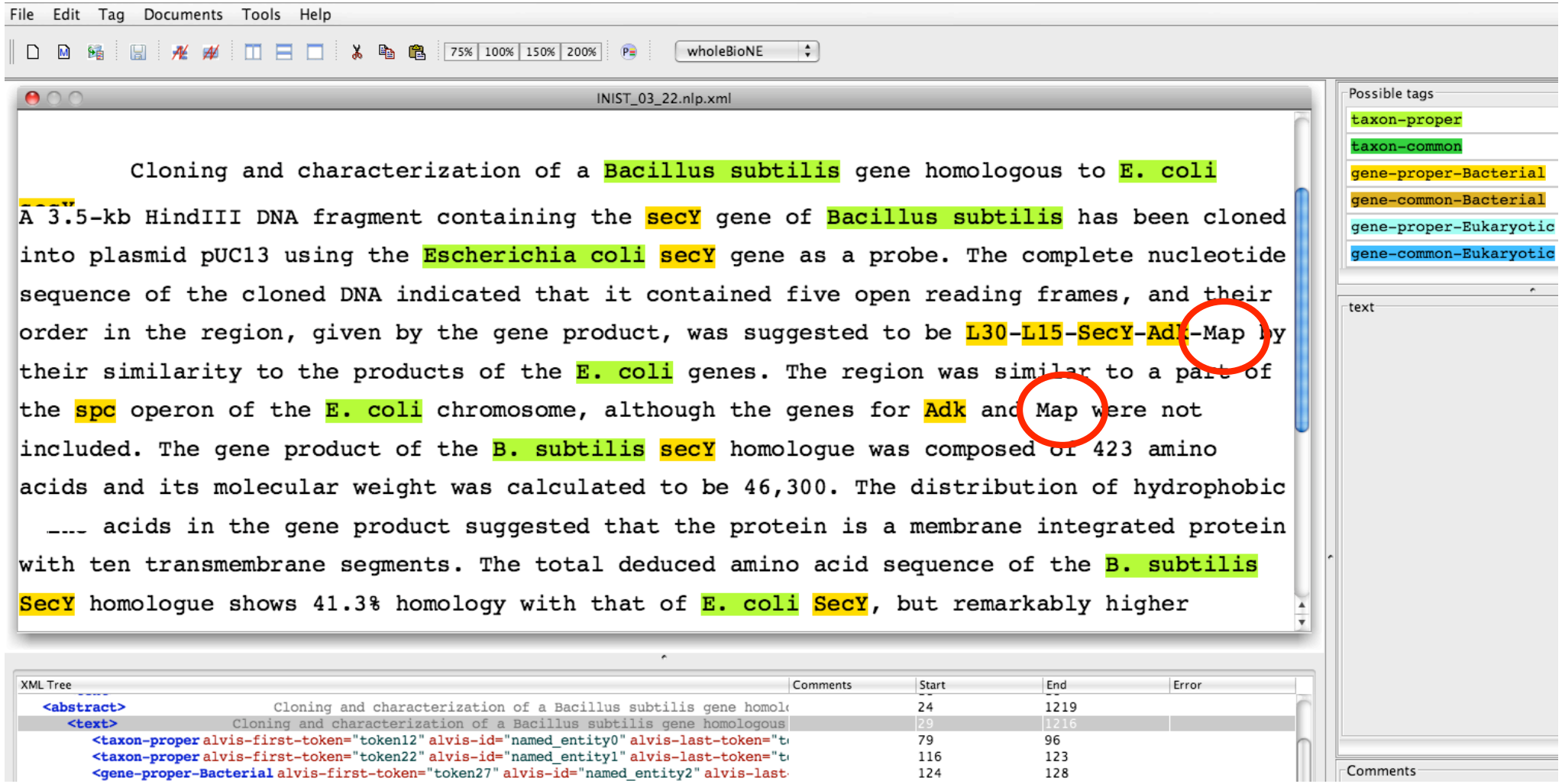
Pattern learning for named entity recognition

Preparation of the learning corpus (one corpus par named entity category)

1. If possible, preannotates the learning corpus by an existing dictionary (list of the names of the category).
2. The names belonging to the category are manually annotated. They represent the **positive examples** (ex. ***GerE*** *protein inhibits transcription in vitro of the* ***sigK*** *gene*).
3. The **negative examples** are automatically deduced from the positive examples: for example, the terms of at most 3 nouns are tagged as as *not NE* (*ex. coat protein, lytic enzymes*)
4. Builds a **relevant representation** of the examples (*ex. nb letters, digits, words in the neighborhood*).

Automatically learning discriminant attributes with a ML method (SVM, ME, C4.5): occurring in positive examples but not in any negative example.

Manual annotation of names for training



Representation of learning examples for NER

- Text segmented into sentences and words
- Structure of the document (candidate NE occurring in the title, in the abstract)

- Typographic information (case, symbols, length, …)
  - **First_upper**: the example is capitalized (^[A-Z])
  - **Middle_upper**: the example contains a non-initial uppercase letter (^.+[A-Z])
  - **Only_upper**: all letters of the example are uppercase (^[A-Z]*$)
  - **Last_digit**: the last character of the example is a digit ([0-9]$)
- Belongs to the dictionary or not.
- Lemma, syntactic category of the example and of the context words.

- Attributes specific to the domain (*ex: trigger words in the NE candidate context*)

  Before: *RNAse accumulate bacterial call collision contrary electrophoretic*

  After: *Pho activate activation analysis bind box dependent domain*

Named entities disambiguation

What is ***ara***? a *bird*, a *promoter*, an *operon? Or a Japanese biologist ...*

Introduction of marker-free deletions in *Bacillus subtilis* using the AraR repressor and the *ara* promoter

Shenghao Liu[1], Keiji Endo[1], Katsutoshi Ara[1], Katsuya Ozaki[1] and Naotake Ogasawara[2]

The *araR* gene encodes the repressor for the arabinose operon (***ara***) of *B. subtilis*. The counter-selective marker cassette consists of a promoterless neomycin (Nm)-resistance gene fused to the ***ara*** promoter [..]

Given `1.[ara = arabinose operon]` is an `operon` `2. [ara]` is a `bird`

Hypothesis The category of first occurrences of ambiguous or new terms is easier to identify by definitory contexts.

Learning method

- For the 1st occ: learn rules *as usual* by supervised classification applied on annotated training corpus
- For the others: include as attribute, the category assigned to the 1st occurrence.

NER disambiguation rules

- The learned rule applied to the *first* occurrence [for the arabinose operon (***ara***) of *B. subtilis*]

If the *candidate*
- Belongs to the *gene* dictionary and,
- At position -2, there is a *gene type* (e.g. *operon*) and,
- At position -1 there is a *comma* or *a parenthesis* (appositive),

Then the *candidate* belongs to the *gene* category.

Evaluation

+ 3,7 recall
+ 1,9 precision

- The learned rule applied to the other *candidate* occurrences. [fused to the ***ara*** promoter]

If the *candidate*
- Belongs to the *gene* dictionary and,
- The first occurrence belongs to the *gene* category,

Then the *candidate* belongs to the *gene* category.

To be improved: Occurrences of different meanings. The 1st occurrence, not the most certain.

Open Question: Relating gene and protein to the right species

The sequences and biochemical functions of many *B. subtilis* and *E. coli* genes are **homologuous**
Their **names are the same** (recA, secA, …)

However,

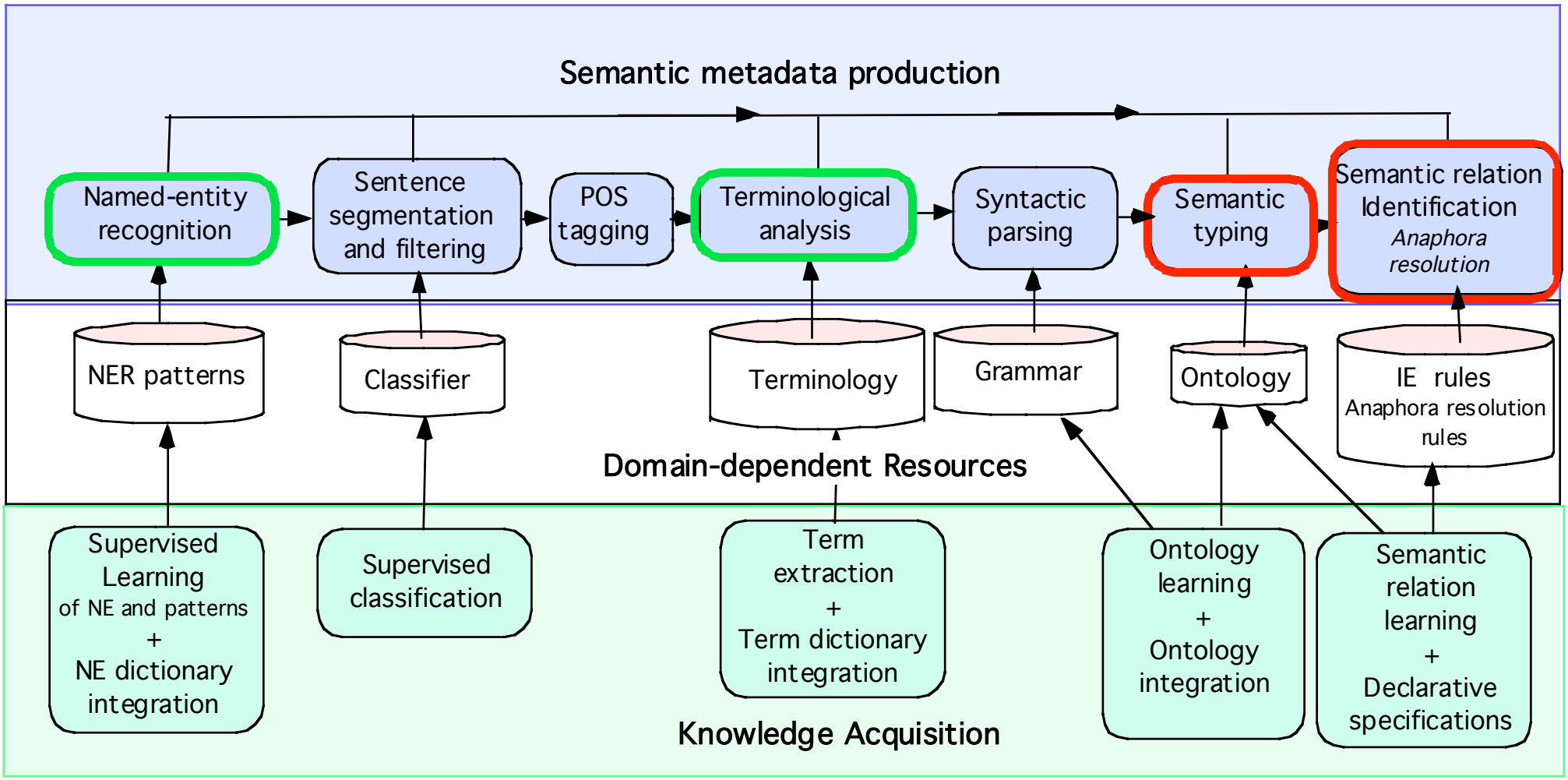
- Their **regulators may differ**
- Exs: Trytophan synthesis. Carbon metabolism (glucose effect)

⇨ **PubMed search for these gene names cannot be restricted to a given species and retrieves many irrelevant papers!**

Attachment of the gene/protein name to the right species could be automatically done

- With Gene-Species attachment rules
- Gene-Species attachment rules can be *reused* for attachment of genes of *any* species
- Simple rules can be *written by hand*
- Complex cases required disambiguating rules to be acquired from manually annotated corpus

BioAlvis, a flexible architecture for *text-mining* in Biology



Identification of relevant keywords: terms

Useful in information retrieval for overcome variation problem and assisting query refinement
First step of design of ontology from corpus

Two steps

1. Acquisition of candidate terms and synonyms with *YateA + FASTR*
2. Expertise of biologists and terminologues needed for

1. The validation of the candidate terms (70%precision)

Too general? *blot analysis*
Too specific? *sigmaA-dependent*
Incomplete? *inverted repeat* *inverted repeat of chaperon expression* *inverted repeat region*

2. The qualification of the relation between terms(synonymy? hyperonymy?)

development of *genetic* competence = development of competence

effect of *low* temperature = effect of temperature

Example of a term extraction tool: YaTea (LIPN)

YaTea combines the use of existing terminologies for achieving a good precision and extraction from training corpus for a good recall.

Input

Corpus tagged by syntactic categories (TreeTagger) and existing terminology.

During[ADV] sporulation[NOUN] of[PREP] Bacillus subtilis[P-NOUN], spore[NOUN]

Method

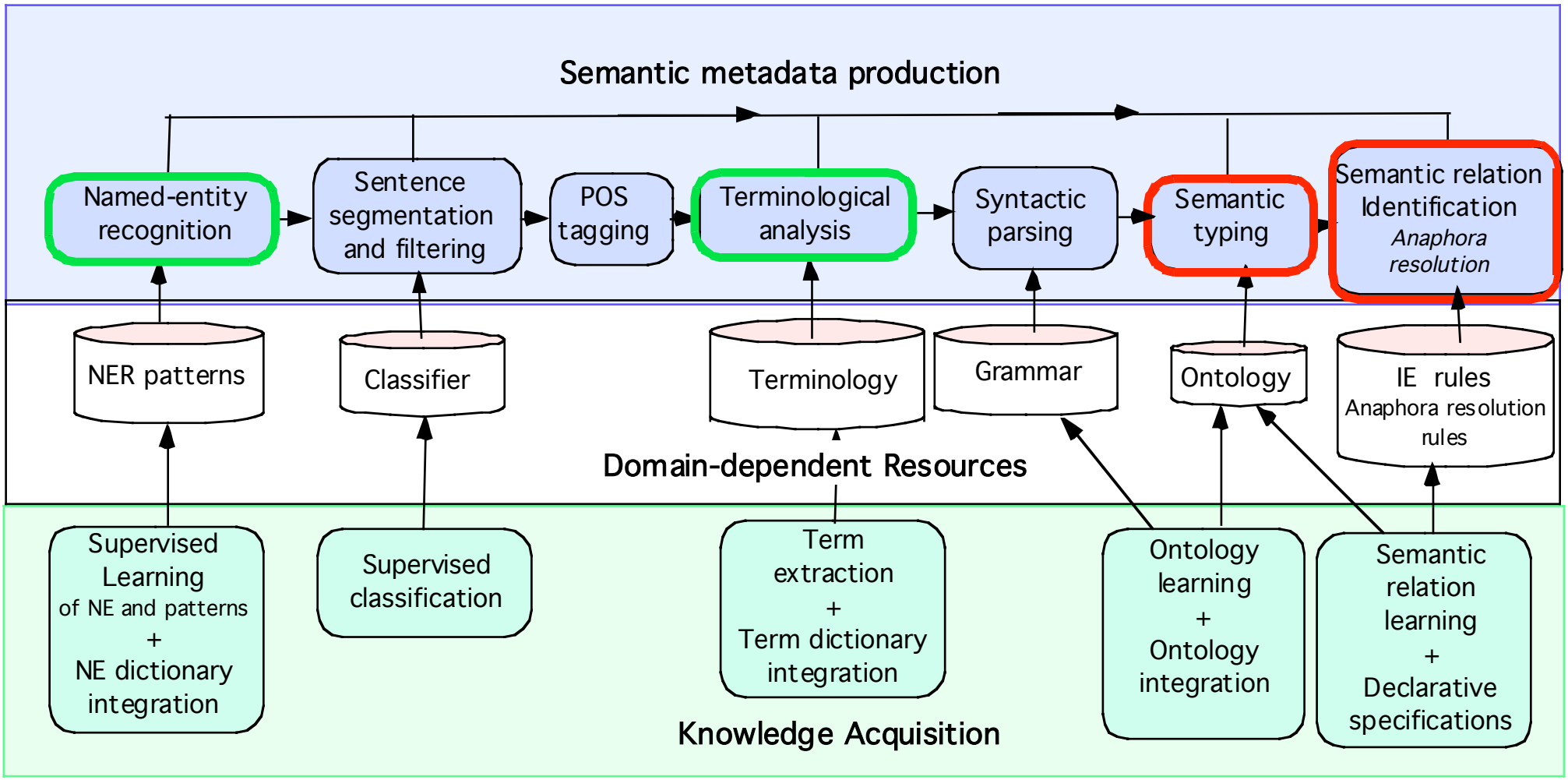
1. « Chunking » based on term border detection.

   During / sporulation of Bacillus subtilis /, / spore coat proteins / encoded by /

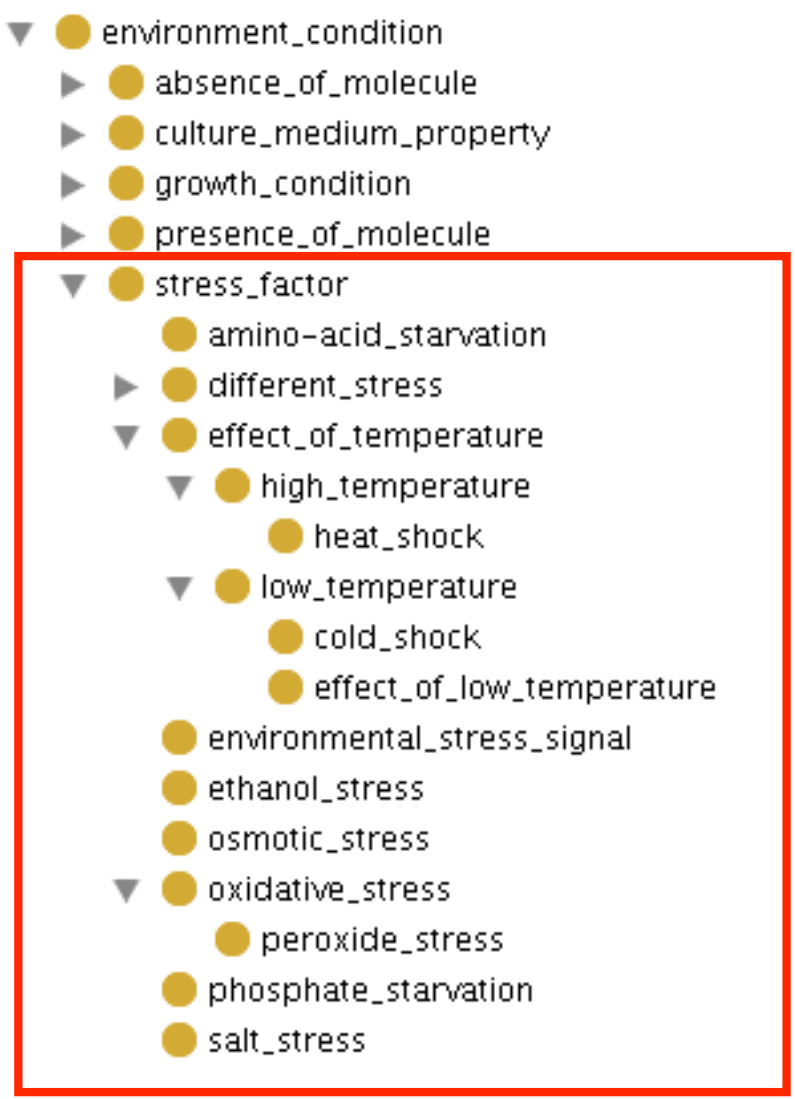
2. Recursive analysis of the « chunks » taking into account,
   - Syntactic patterns NOUN NOUN
   - Forbidden structures and components (*of course*)
   - Patterns specific to certified terms (*in vitro)*

3. Manual validation of the results

BioAlvis, a flexible architecture for *text-mining* in Biology



Term hierarchies for concept indexing



4,500 concepts

Machine learning of conceptual hierarchies from corpus

Goal

- Learning concepts defined as semantic classes of semantic units (terms / NE).
- Learning the Is-a relation between concepts (inclusion des classes).

Two main complementary classes of methods

- **Distributional semantics** (conceptual clustering): easy to implement, very productive, difficult to control and evaluate.
- **Learning Information Extraction rules**: requires annotated examples, but easy to control and evaluate.

In both case, the more homogeneous and normalized is the corpus, the better is the quality of the relations learned.

Acquisition of hyperonymy relations by IE patterns

Based on

- The **internal structure of terms** (compositionality)

  gamma-irradiation stress → gamma-irradiation stress IS-A **irradiation stress**

- **IE patterns** (Hearst')

  Examplification *B. anthracis* ***is an example of*** *a recently emerged* ***pathogen***
  → B. anthracis IS-A **pathogen**

  Enumeration ***spore coat proteins, including*** *CotA*, *CotB*, ***and*** *CotF*
  → CotA IS-A **spore coat protein**

Patterns are more or less productive and reliables

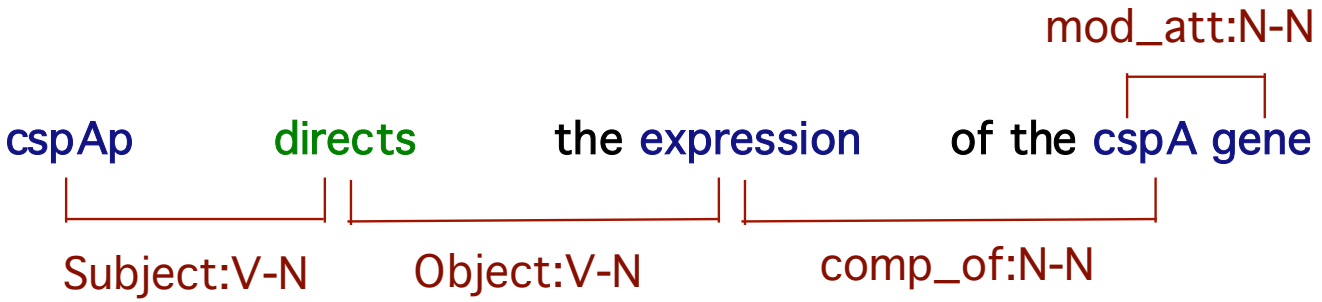
The target ***appears to be one of*** *the* ***proteins***

The interpretation of the relation as a property or as a category is a modeling issue

Bacillus subtilis ***is a*** *strict* ***aerobe***

Learning hyperonymy relations by distributional semantics

- **Syntactic dependencies of the corpus** between heads and arguments → **Learning examples** with occurrence frequency (Asium [Faure & Nédellec, 1999])



synthesis ***Comp:N-N(during)*** ***Term***

[synthesis] [**during** growth].

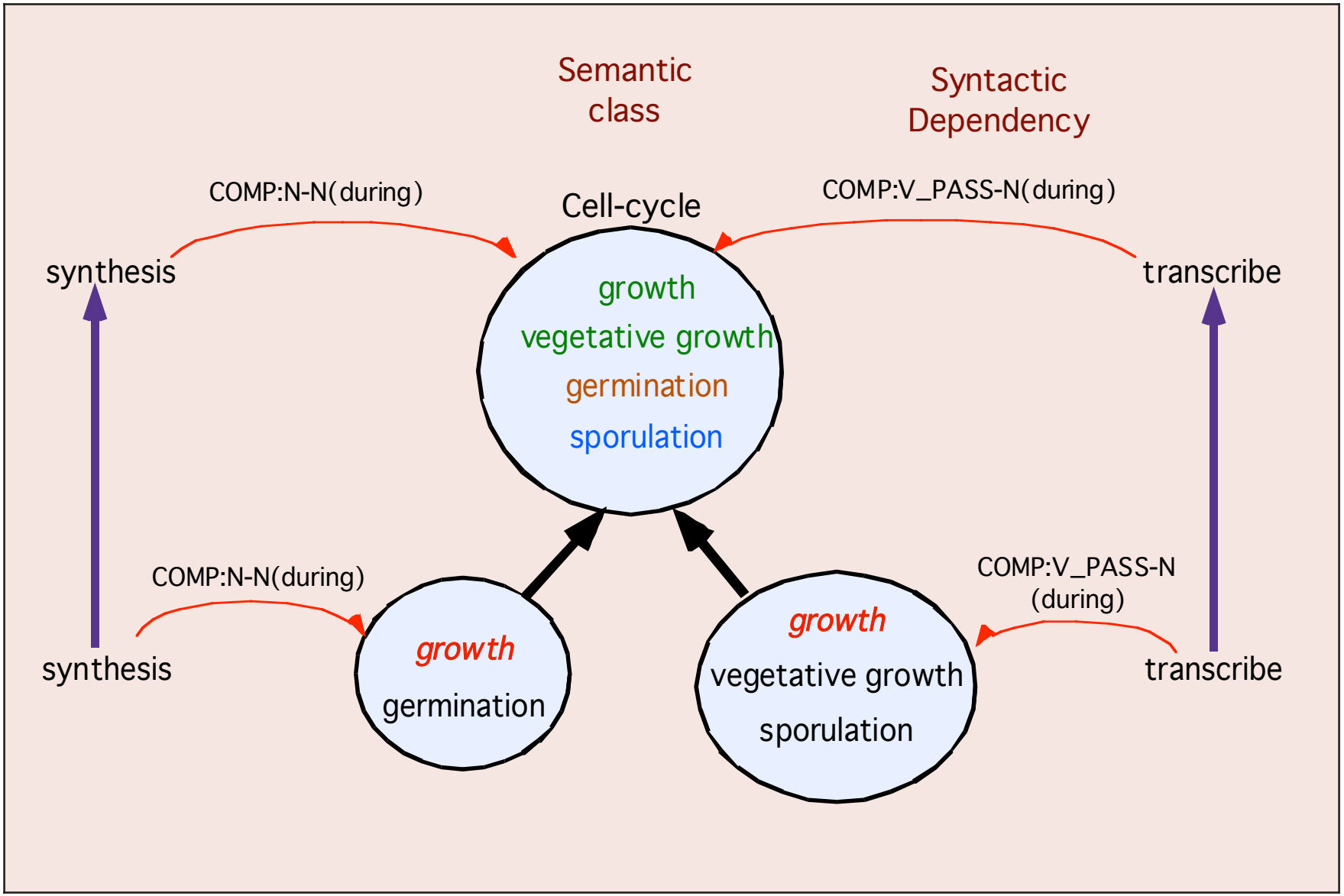
[synthesis] [**during** germination].

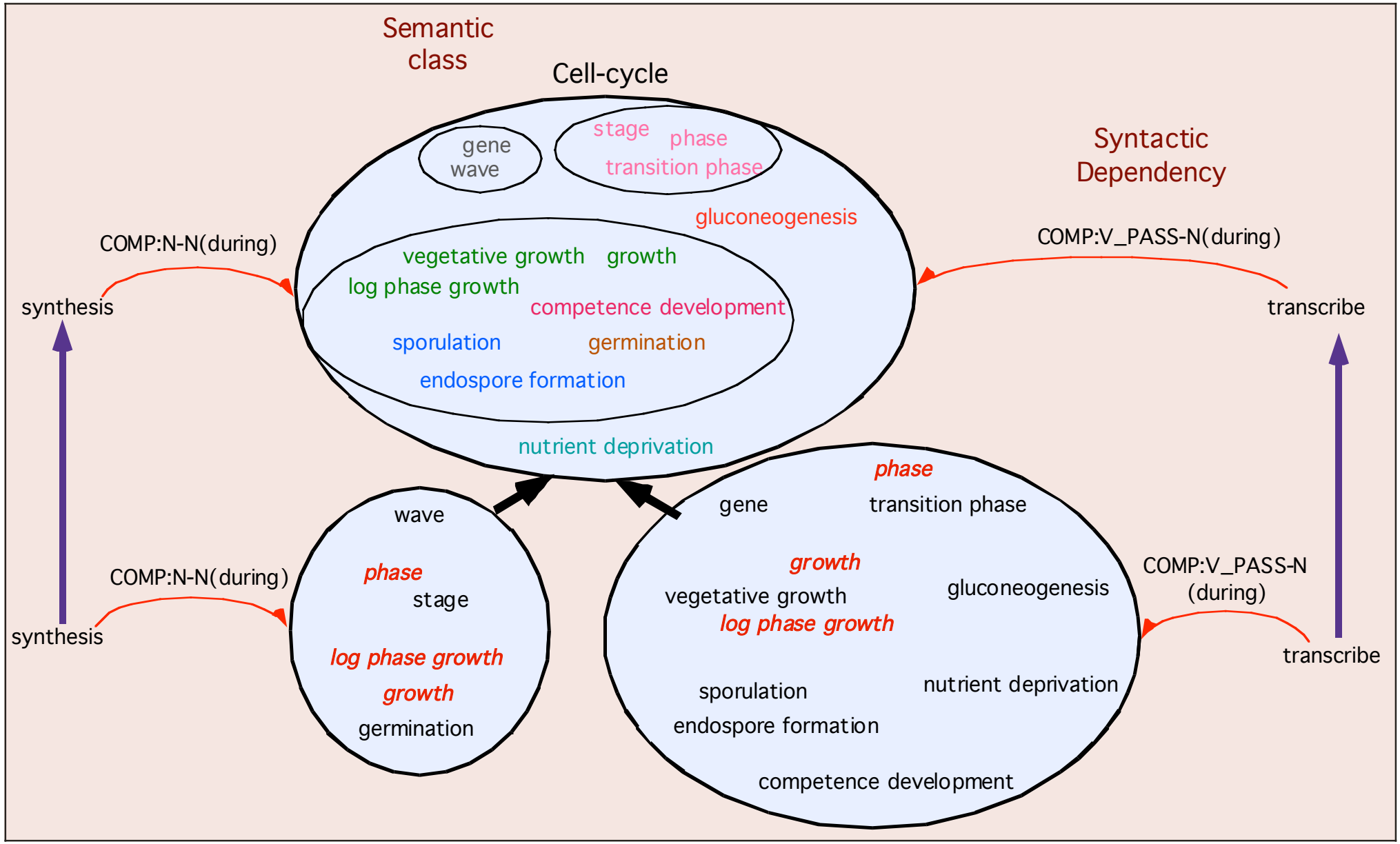
transcribe ***Comp :V_Pass-N(during)*** ***Term***

[transcribe] [**during** growth].

[transcribe] [**during** vegetative growth].

[transcribe] [**during** sporulation].


Semantic class
Syntactic Dependency
COMP:N-N(during)
Cell-cycle
COMP:V_PASS-N(during)
synthesis
transcribe
growth
vegetative growth
germination
sporulation
COMP:N-N(during)
COMP:V_PASS-N (during)
synthesis
growth
germination
growth
vegetative growth
sporulation
transcribe

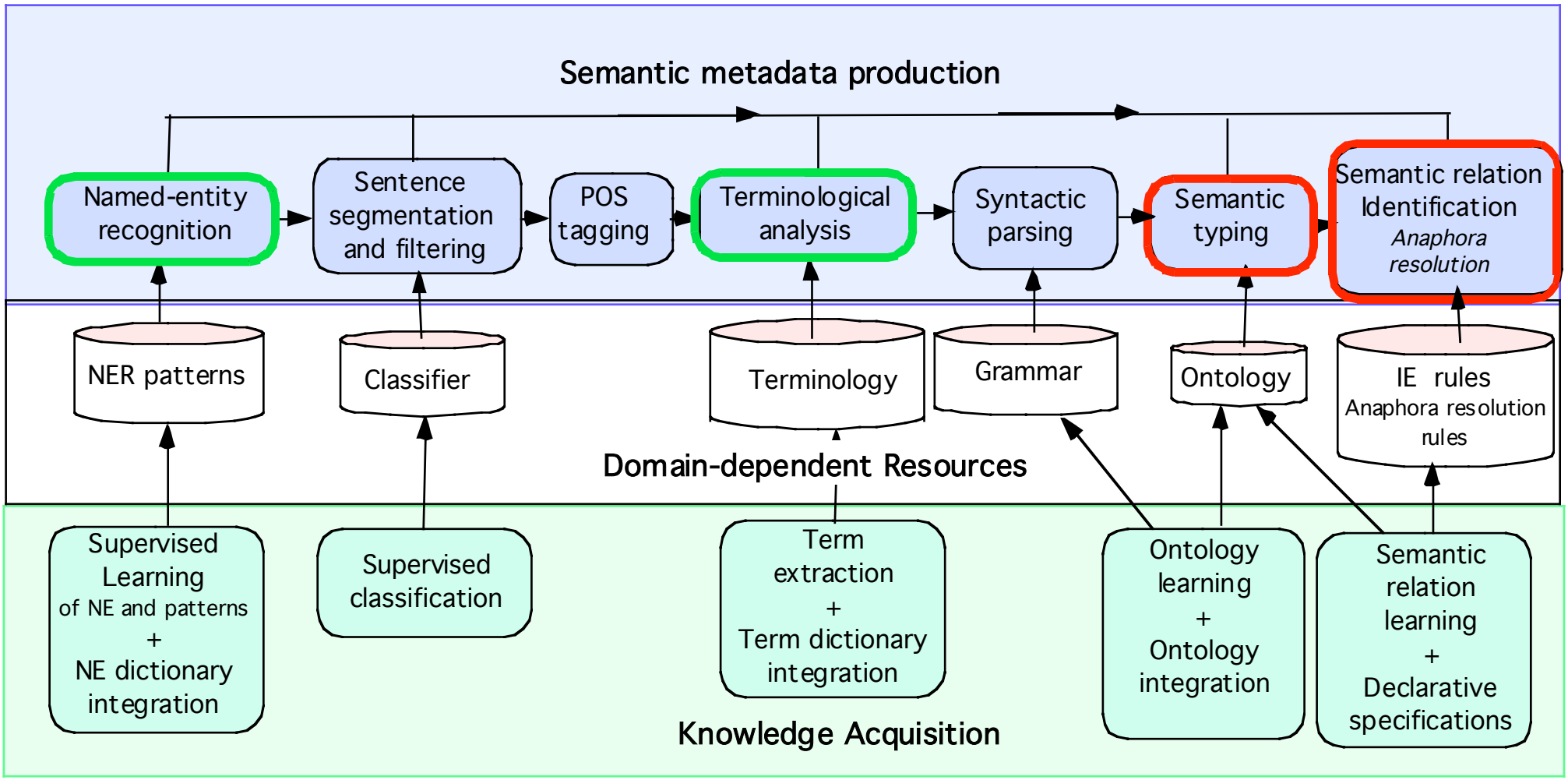

Semantic class
Cell-cycle
gene
wave
stage
phase
transition phase
gluconeogenesis
vegetative growth
growth
log phase growth
competence development
sporulation
germination
endospore formation
nutrient deprivation
Syntactic Dependency
COMP:N-N(during)
synthesis
COMP:V_PASS-N(during)
transcribe
wave
phase
stage
log phase growth
growth
germination
gene
transition phase
growth
vegetative growth
gluconeogenesis
log phase growth
sporulation
nutrient deprivation
endospore formation
competence development
COMP:N-N(during)
synthesis
COMP:V_PASS-N (during)
transcribe

Combining distributional semantics and patterns

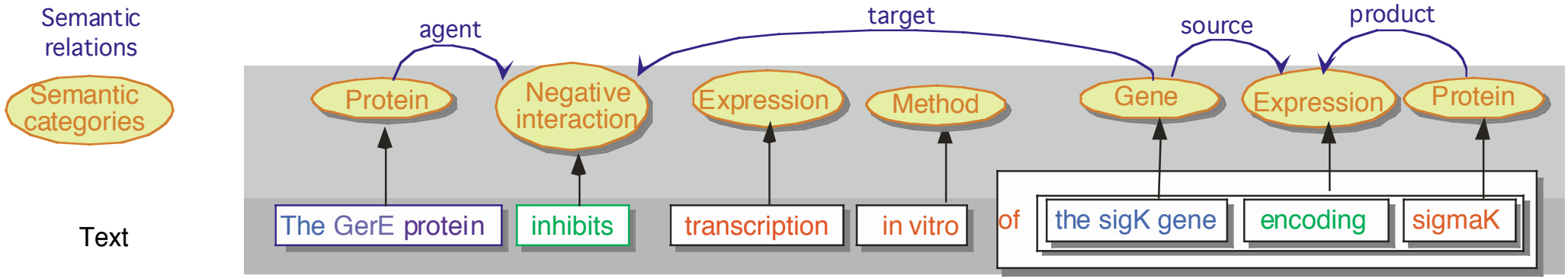
Semantic classes are split along clustering with respect to the hyperonyms identified by IE patterns.

Examples from Web pages	*Patterns*
Gluconeogenesis (abreviated **GNG**) is a metabolic pathway Pathways for Glucose Synthesis (**Gluconeogenesis**) **Gluconeogenesis** refers to a metabolic pathway that synthesizes a glucose molecule	is a *apposition* refers to
stresses such as radiation, temperature or **osmotic shock**, **oxidative stress** and **nutrient deprivation** many stress adaptations, including **osmotic**, **oxidative**, **desiccation**, **carbon**, and **nitrogen stress** stresses, such as high levels of UV light, gamma radiation, **heat**, **pressure** and **desiccation**	such as *enum* including *enum* such as *enum*
The first phase of growth is the **log phase,** The second phase of growth is the **logarithmic phase** (**log phase**), also known as the **exponential phase** The final phase of growth is the **stationary phase** divided into two different phases, i.e., **vegetative growth phase** and **development phase**.	is a is a is the i.e., *enum*

BioAlvis, a flexible architecture for *text-mining* in Biology

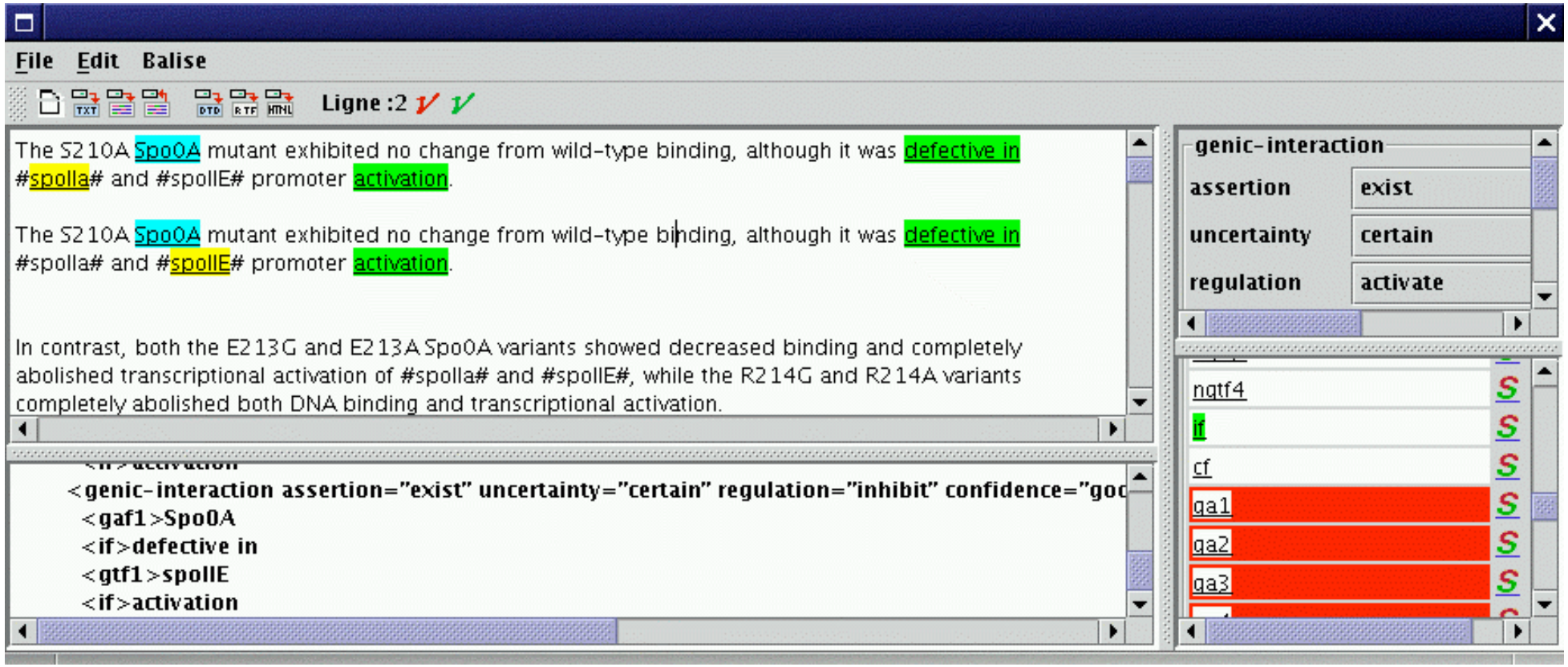


Tagging semantic relations



Interaction	**Type :** negative	
	Agent : GerE protein	
	Target : **Expression**	**Source :** sigK gene
		Product : sigmaK

Extraction rules to be acquired from training corpus



*See **LLL dataset** "Learning Language in Logic" challenge (ICML05)*

Example of results on relation learning

- Training data: gene interactions (agent, target) in *Bacillus subtilis*
  LLL international challenge dataset on "action without coreference"

- Linguistic normalization and abstraction

- Rule learning with **LP-Propal** [Manine et al., 2008]

	F-mesure
[Goadrich et al., 2005], without linguistic information	58,5
[Riedel and Klein, 2005] with linguistic information	65,5
[LP-Propal] with linguistic information	89,4

Conclusion

Text-mining is not a method, but a large panel of methods and technologies

Various technologies are integrated in operational on-line services
They rely on domain specific knowledge bases

Resarch in Machine Learning and Natural Language Processing improves

- the quality of the information extraction
- the automatization of the acquisition of knowledge

Cooperation between biologists and computer scientists is critical for defining the needs and qualifying the knowledge to be acquired (non-trivial)!

Learning extraction rules

Learning method

Supervised relational learning,
Horn clauses
Multi-class learning: top-down ILP method Propal [Alphonse, 2003]

Training data pre-processing

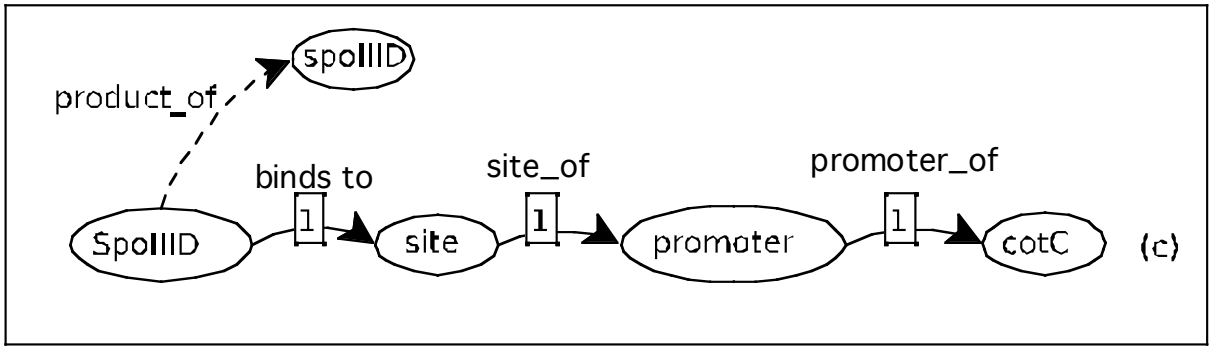
1. Selection of relevant documents.
2. Segmentation and filtering of relevant sentences.
3. Manual annotation of the relations in the positive training data.
4. Negative example generation (near-miss selection in relevant sentences under closed-word assumption)
5. Training example preprocessing (linguistic processing and saturation by BK).

Application of the learning method for acquiring the rules representing the discriminant linguistic attributes.

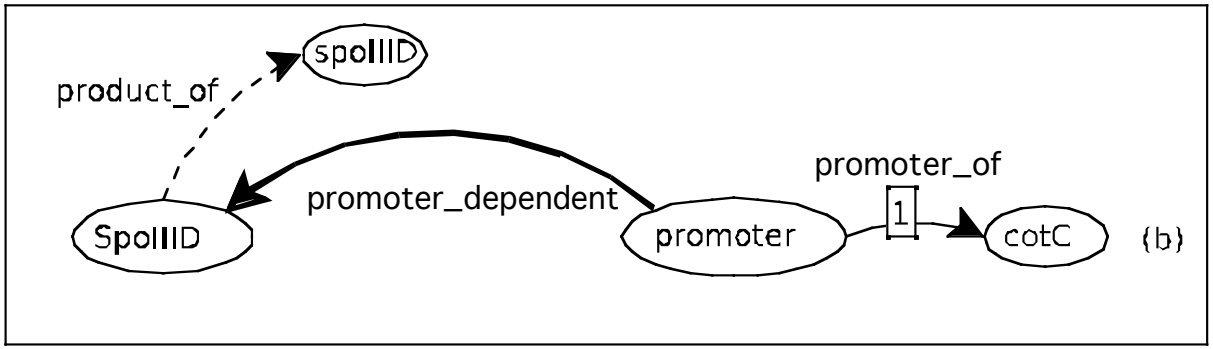
Sentences in natural language

SpoIIID binds strongly to two sites in the cotC promoter region.

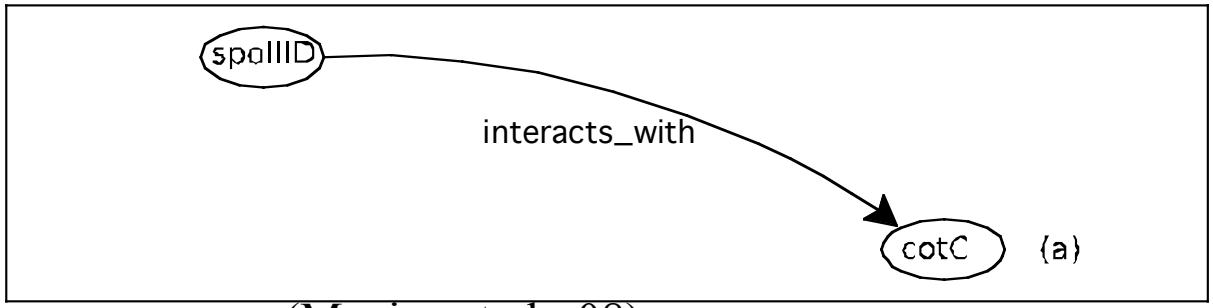
Syntactic and semantic analysis



Derivation of the relation promoter-dependent

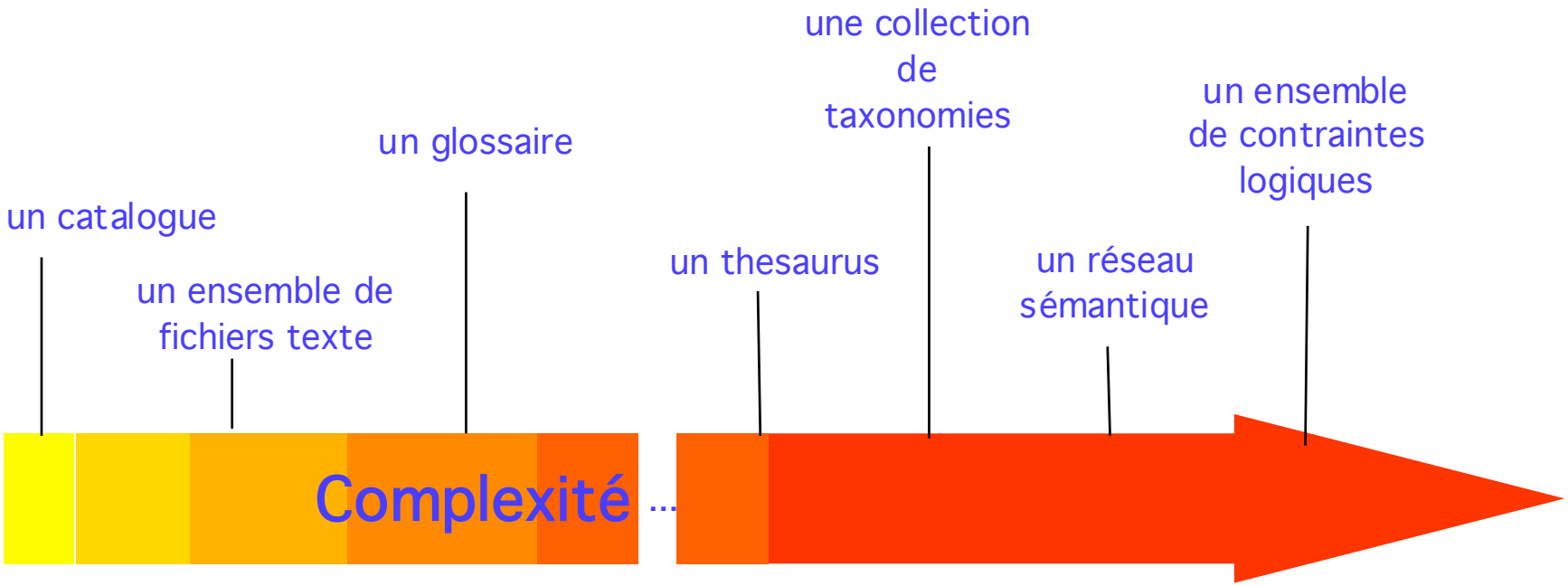


Derivation of the interaction relation



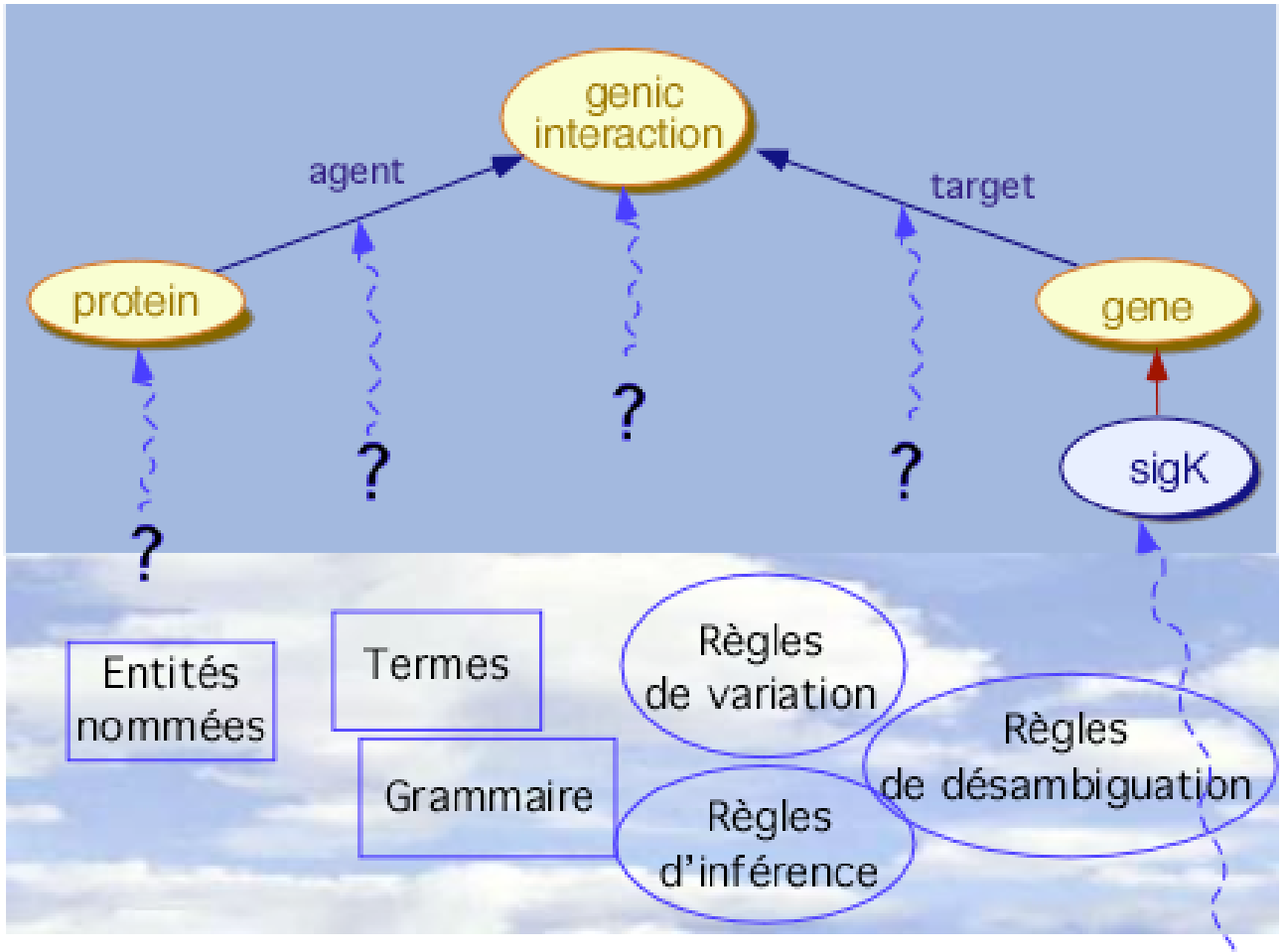
(Manine et al., 08)

Une ontologie est ...

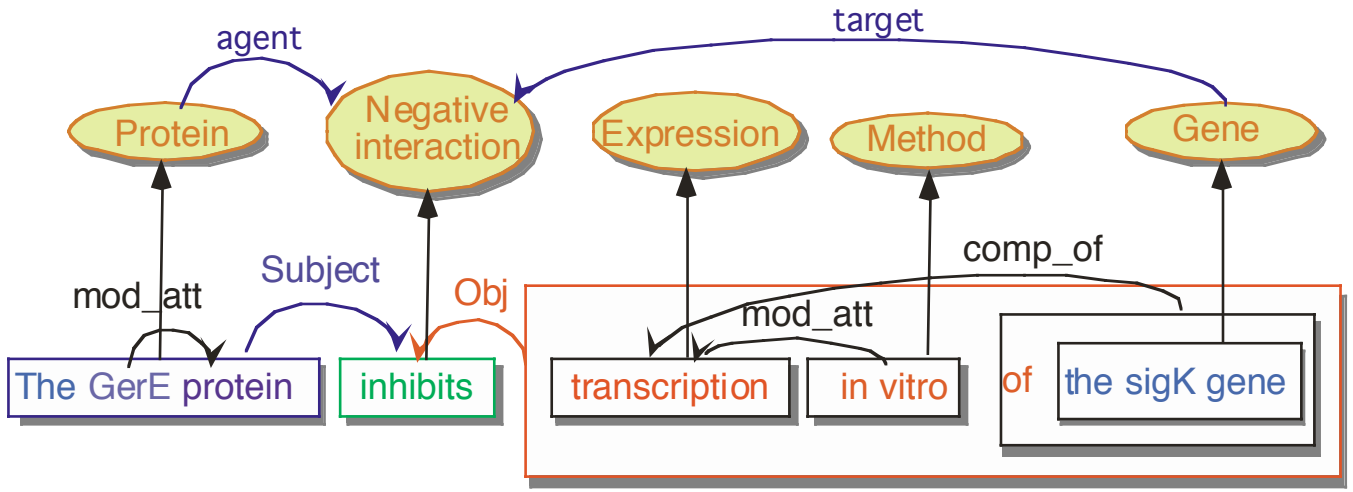


Nature du lien formel avec le texte ?

Interprétation incomplète

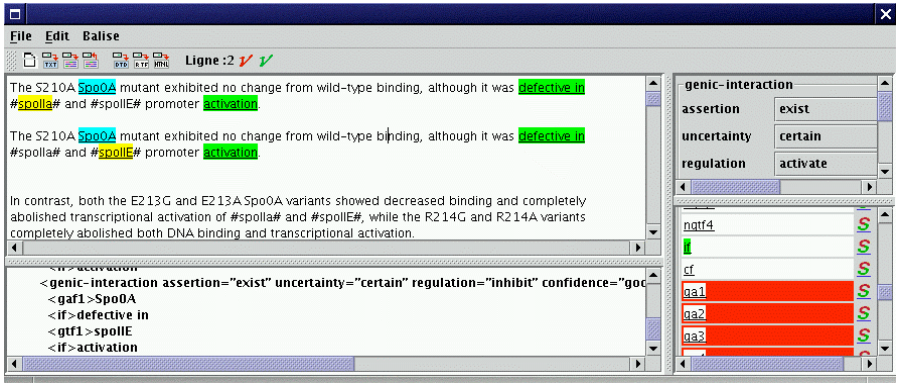


SpoIIID product is required for the transcription of sigK



Relation sémantique	agent(Ger_protein, inhibit), target(sigK gene, inhibit), …
Catégorie sémantique	is_a(Ger_protein, protein), is_a(inhibit, negative_interaction), …
Dépendance syntaxique	sujet(Ger_protein, inhibit), obj(transcription, inhibit), …
Terminologie	terme(GerE protein), terme(in vitro) terme(sigK gene)
Catégorie grammaticale	cat(the, dét), cat(GerE, nom), cat(inhibit, verbe), …
Entité nommée	entité(GerE, protein), entité(sigK, gene), entité(sigma K, protein)
Texte segmenté	mot(the), mot(Ger_protein), mot(inhibit), mot(transcription), …

Various représentation of learning examples



/Combined /action /of /two /transcription /factors /regulates /genes /encoding /spor
/proteins /of /Bacillus /subtilis /. /During /sporulation /of /Bacillus /subtilis /, /spore /
/proteins /encoded /by /cot /genes /are /expressed /in /the /mother /cell /and /deposited
/the /forespore /. /Transcription /of /the /cotB /, /cotC /, /and /cotX /genes /by /final /
/K /) /RNA /polymerase /is /activated /by /a /small /, /DNA - /binding /protein /called
/. /The /promoter /region /of /each /of /these /genes /has /two /GerE /binding /sites /.