Introduction to Natural Language Processing

INTRODUCTION:

Towards industrial applications of computational linguistics:

Corpus based-linguistics

Linguistic Processing Levels

Martin Rajman

Martin.Rajman@epfl.ch

and

Jean-Cédric Chappelier

Jean-Cedric.Chappelier@epfl.ch

Artificial Intelligence Laboratory

About the course

WebSite: `coling.epfl.ch/`

GRADING:

- 4 quiz during semester 25% (i.e. 6.25% each):

  45 minutes each.

  ☞ see the website for the dates

- final exam: 75%

  3 hours.

Objectives of this lecture

- Introduce natural language and its functions
- Show possible applications/realizations and associated constraints

Contents

- NLP Applications
- Functions of Natural Language
- Industrial Constraints
- Corpus-Based Approach to NLP

Natural Language Processing/Understanding

Natural Language Processing is (and has long been) a great challenge in AI:

- How can we construct a comptuter representation (and which one?) from a observed text?
- How can we generate (natural) text from computer representations?

We don't know yet how to properly model human language (nor thoughts).

We instead rely on learning from data and performance on specific tasks.

Modeling the task(s) can still lead to new insights on the origin.

Main Application Domains

- *Automated Translation*
  - Second World War, European Community, Canada, Switzerland, ...

    (Systran, Reverso, ...)
- *Writing Assistance*
  - Spelling error correction (Cordial, Ispell, MS-Word, ...)
  - Text generation (Canadian weather forecast, Financial reports, ...)
  - Summarization tools
- *Information Retrieval / Web Search / Information Extraction* (Google, ...)
- *Information filtering and classification*
  - emails, news, patents, ...
- *Natural Language Interaction / Interfaces*
  - Vocal Command
  - Vocal Access/Servers (phone-book inquiry, ...)

Natural Language

Natural .vs. **Formal**:

- formal languages are by construction **explicit** and **non-ambiguous**
- natural languages are in essence **implicit** and **ambiguous**

implicit:

Remove the stones from the cherries and put them in the pie.

The hunter shot the tiger; his wife too.

ambiguous:

Time flies like an arrow.

She was eating a fish with { *bones.* / *anger.* / *some friends.* / *a fork.*

Natural language functions

COMMUNICATION:

✏ Conciseness

The student gave his homework to the professor who told him that it could have been better.

The student gave the homework of the student to the professor. The professor told the student that the homework of the student could have been better.

```
Student.give(Student.homework, Professor);
Professor.tell(Student, be_better(Student.homework));
```

✏ Shared knowledge

– I gave him a nice pen.

– A "Mont Blanc"?
– Yes, this brand is really great!

– How large is it?
– Well, big enough for 20 head of cattle.

Natural language functions (2)

REPRESENTATION:

☞ unlimited **expressive power**

logical expressions of any order:

Earth is curved.	`curved_earth = TRUE`
All politicians lie.	$\forall$ `x, politician(x)` $\Rightarrow$ `lier(x)`
Everything quickly done is not well done.	`quickly(do)` $\Rightarrow$ $\forall$`x, [ do(x)` $\Rightarrow$ `not good(x) ]`

and even non sense!

Following the antagonist bi-polar logic, it could be assumed that we enter a kind of "T-state" in which imaginary/rational-real updating and potentiation tend towards a dynamic stability...

Natural language functions (3)

Why is natural language implicit and ambiguous?

implicit enables conciseness (ellipsis, anaphoric references, ...)

... but entails potential ambiguities

unlimited expressive power requires flexible interpretation rules

... and therefore forbids the meaning to be exclusively expressed by the surface form.

How is it possible that we still understand each other?

☞ very large amount of shared knowledge

(previous) Examples:

- we usually eat cherries and not their stones
- we assume that hunters are not all criminals
- we know that a writing pen is never big enough for 20 head of cattle

NLP and Industrial Applications

For real-world NLP applications:

- **Task specification is essential:** even a small modification of the targeted task may turn a NLP application from feasible to impossible to acheive

  Example: computer assisted translation .vs. automatic translation of free text (e.g. Web)

- **effective usefulness:** NLP is not always the best solution and sometimes other means/medias should be prefered

  ☞ Think about how to **evaluate the "*usefulness* / *drawbacks*"** ratio as well as the ***implementation difficulties*** when planning to introduce NLP into some application.

Constraints due to the application context

The two main application contexts correspond to the two main functions of language:

① language = communication tool

⇨ e.g. applications for interfaces

Constraint: **Real Time**

$\simeq$ 180 word/mn $\longrightarrow$ 1 word every 300 ms

② language = knowledge representation formalism

⇨ e.g. applications for Information Retrieval

Constraint: **huge amounts of data** (to compensate the still relatively poor performence)

10'000 documents within 1 day $\simeq$ 300 word/s $\longrightarrow$ 1 word every 3 ms

Constraints due to the application context (2)

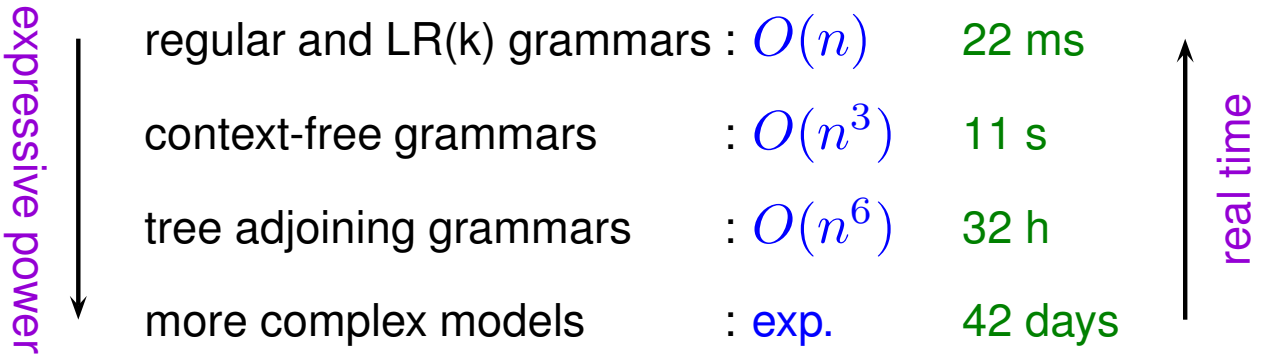
The constraints imposed by real-world NLP applications entail the need for:

1. fast processing $\Rightarrow$ polynomial algorithms
2. a good coverage of the (sub-)language corresponding to the considered application
   $\Rightarrow$ sufficient linguistic resources

Choice of Language Models

The use of polynomial-time algorithms impose severe limitations on the complexity of the linguistic models to be considered

Examples of how algorithimc complexity is related to the expressive power of grammatical models:



$\Rightarrow$ For industrial applications: **context-free grammars** (or various extensions of these)

Why is NLP difficult?

- **lack of linguistic competence**
- power laws (at all levels)
- curse of dimensionality (high dimension + sparseness)
- subjectivity (Inter-Annotator (dis)Agreement)
- multi-scale

Industrial NLP

Linguistics skills are quite **difficult to find** in the industry

and

linguistic resources are often **difficult** (= costly) **to produce**

$\Rightarrow$ Resources may be at least as costly as the design of the core system itself

and this is even more the case for resources at the **semantic level**

$\Rightarrow$ **Industrial** approaches currently considered mainly focus on the **structural analysis** of the sentence (grammar). Techniques dedicated to the **understanding** of the meaning (semantic) are less developed.

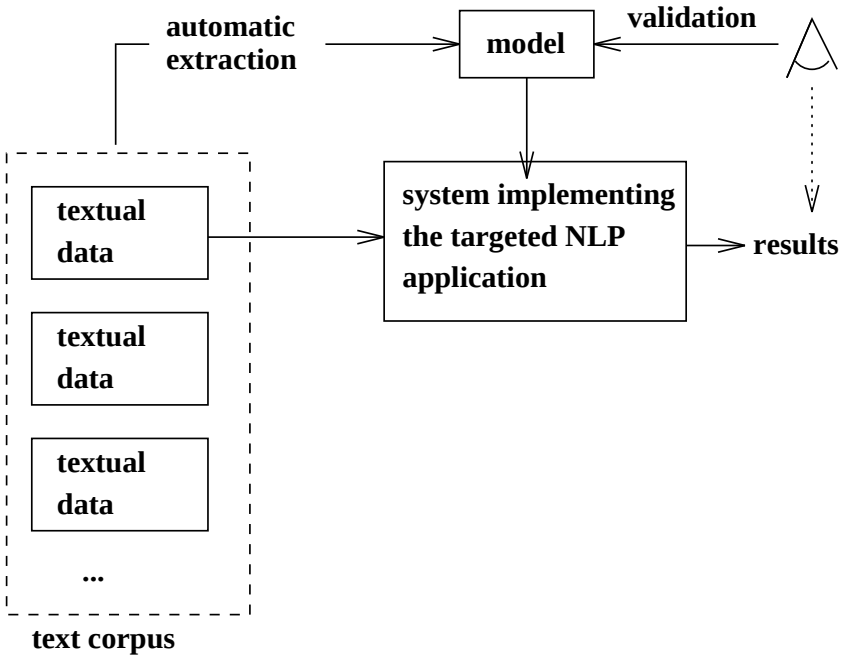
$\Rightarrow$ The integration of semantics/pragmatics is made through the **(automated) extention** of syntactic models (e.g. probabilistic models).

Corpus-based linguistics

The goal is thus not so much to reproduce the human linguistic competence with approaches that try to model our understanding of language, but...

..to reproduce, *for a given task (applicative framework)*, the corresponding linguistic behaviour with models that can be (semi-)automatically trained from large amounts of textual data representative for the considered task.

Corpus-based linguistics (2)



The evaluation of the considered models does not try to measure their explanatory power (about human language) but the improvement of their performance for the considered application

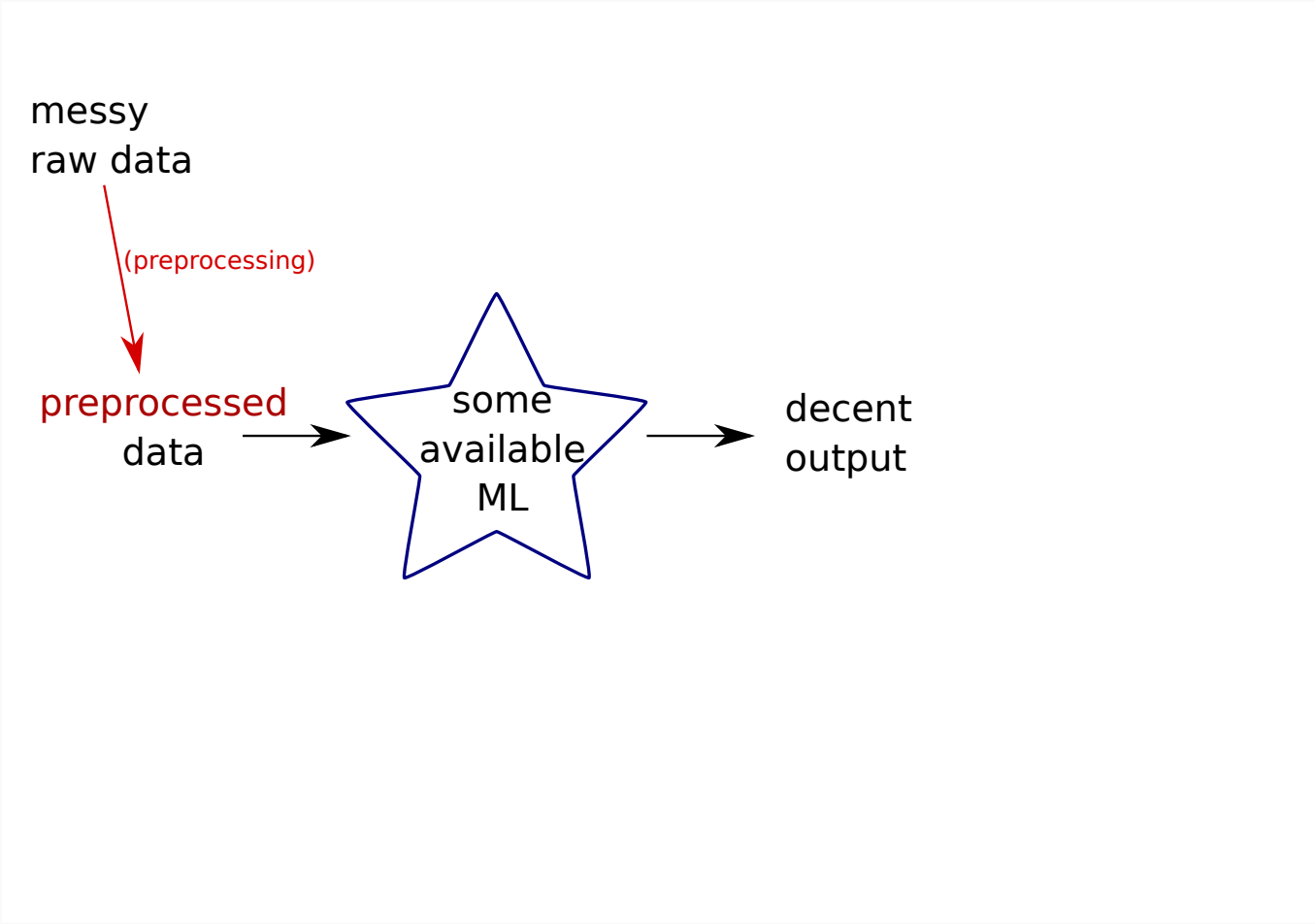
☞ **Corpus-based**, **Performance-oriented** computational linguistics

So, is this course a Machine Learning Course?

Machine Learning:

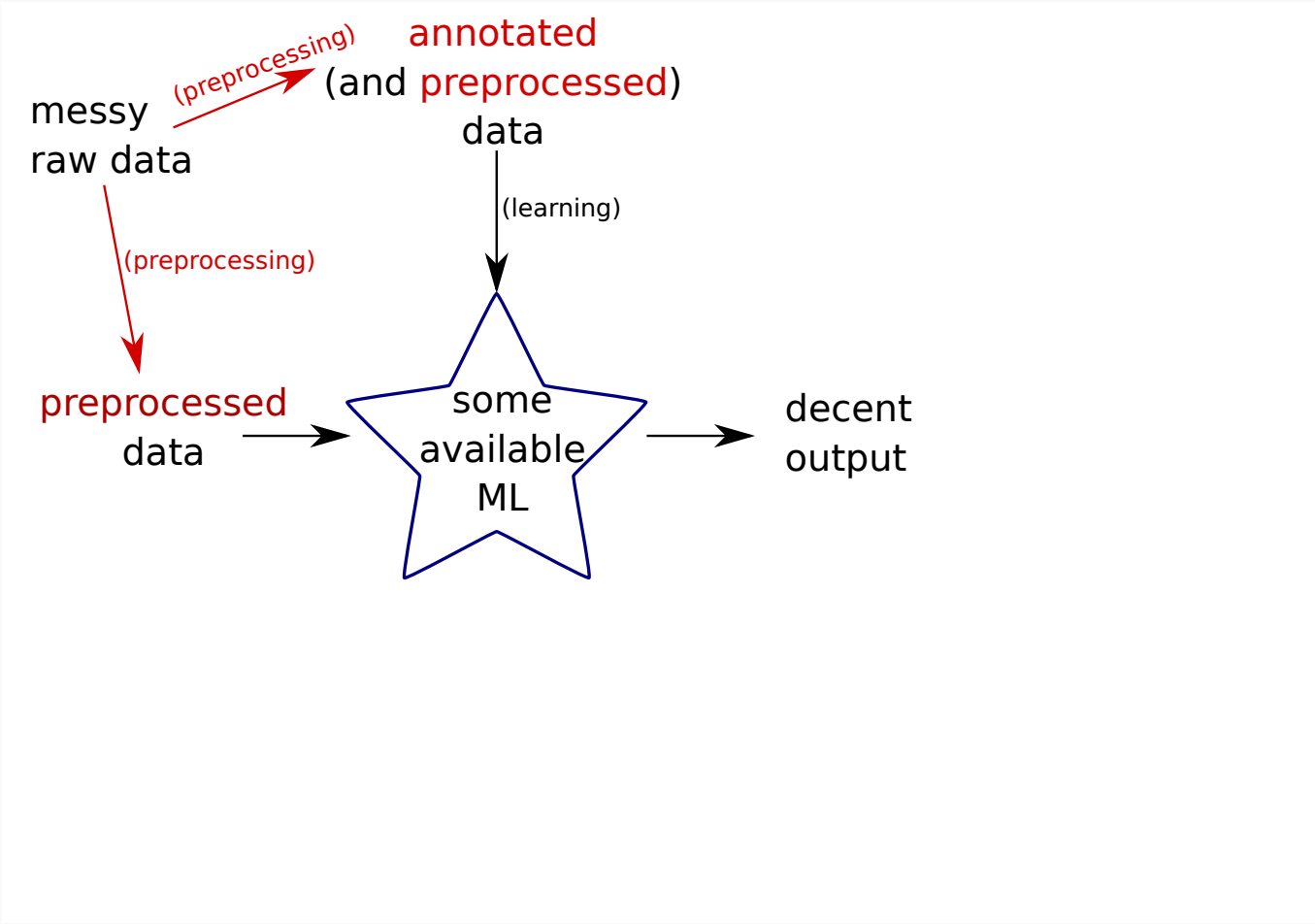
So, is this course a Machine Learning Course?

Machine Learning: good preprocessing is still (very) important



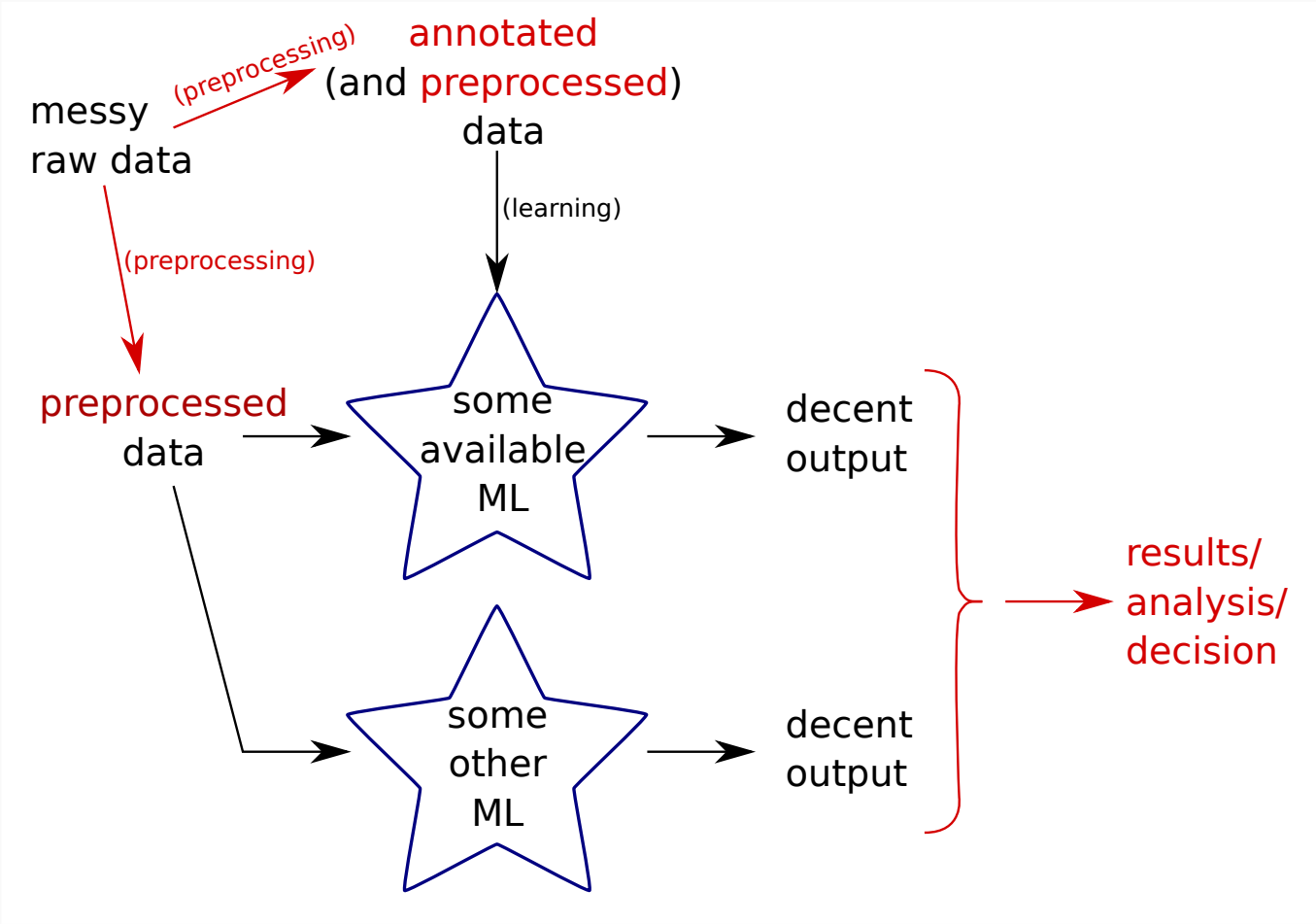
So, is this course a Machine Learning Course?

Machine Learning: need for (good) supervision



So, is this course a Machine Learning Course?

Machine Learning: need to understand (origins of) outputs, analyze errors, ...



So, is this course a Machine Learning Course?

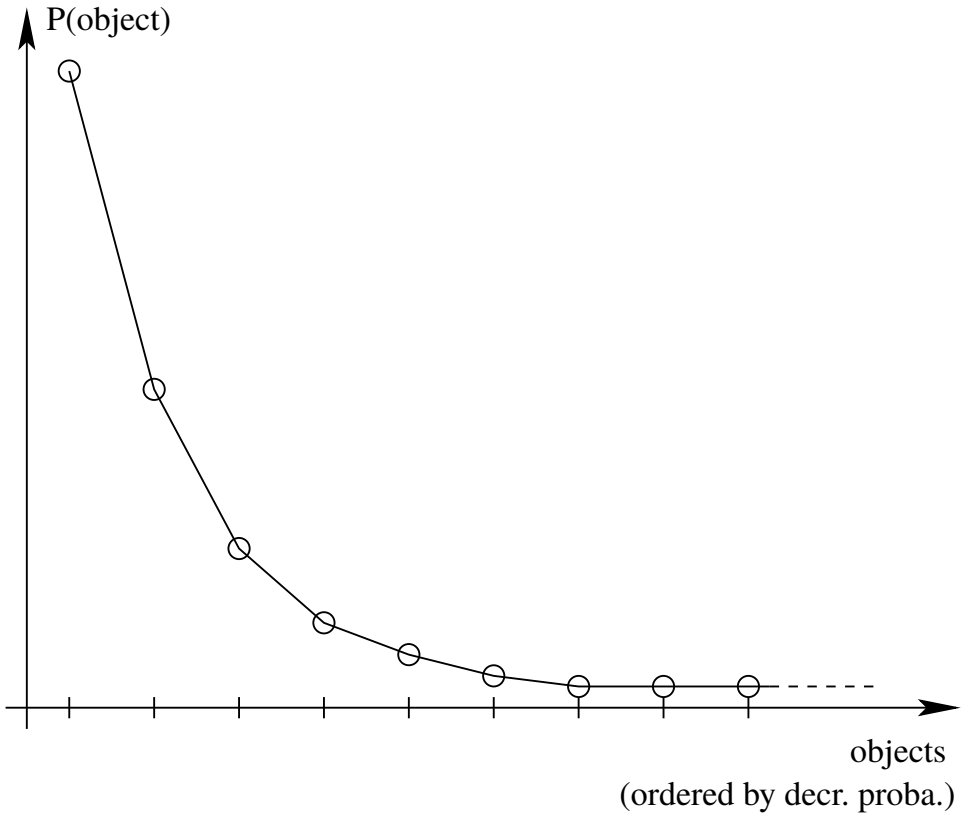
- NLP makes use of Machine Learning (as would Image Processing for instance)
- but good results require:
  - good preprocessing
  - good data (to learn from), relevant annotations
  - good understanding of the pros/cons, features, outputs, results, ...

☞ The goal of this course is to provide you with specific knowledge about NLP.

Why is NLP difficult?

- lack of linguistic competence
- **power laws** (at all levels)
- curse of dimensionality (high dimension + sparseness)
- subjectivity (Inter-Annotator (dis)Agreement)
- multi-scale

The impact of power laws
(e.g. Zipf Law, "Zeta distribution")



Example (Brown Corpus):

most frequent word ("*the*"): $\simeq$ 7% of all word occurrences

(69971 over 1 million)

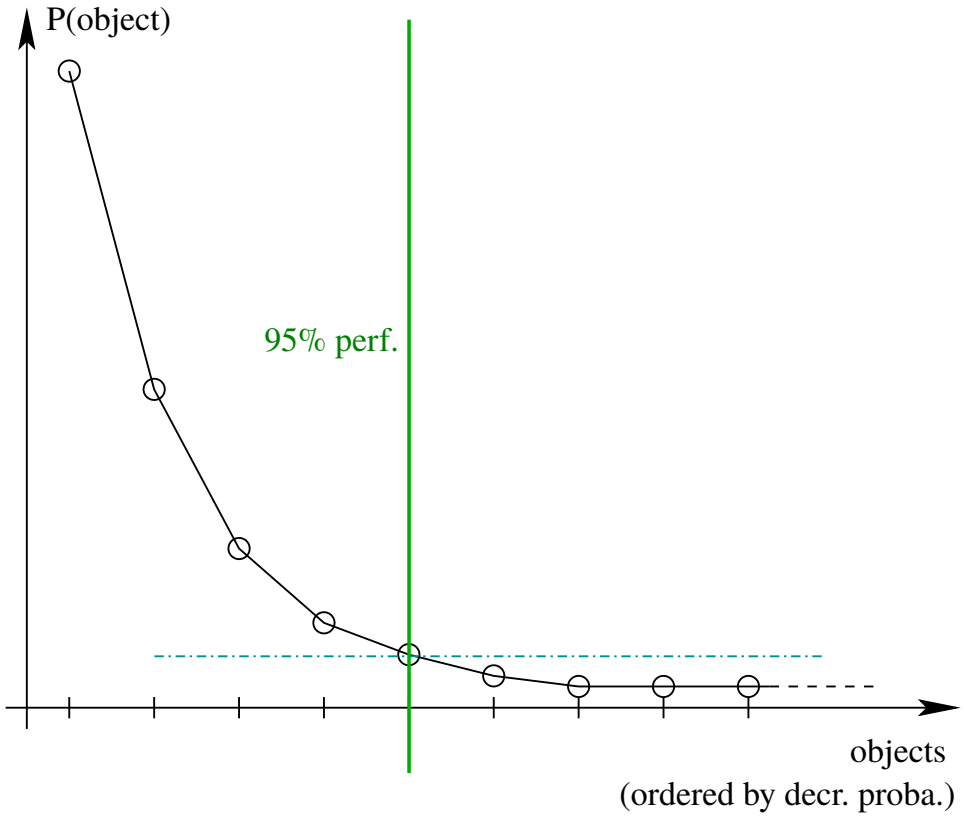
second most frequent ("*of*") : 3.5%

Only 135 different words make 50% of the corpus (occurrences)

Conversely 50% of the vocabulary (not the same % !!) are hapaxes (1 occurrence)

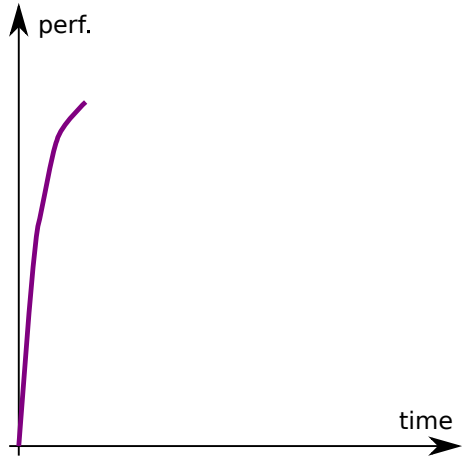
(that cover 2.5% of the corpus)

The impact of power laws
(e.g. Zipf Law, “Zeta distribution”)



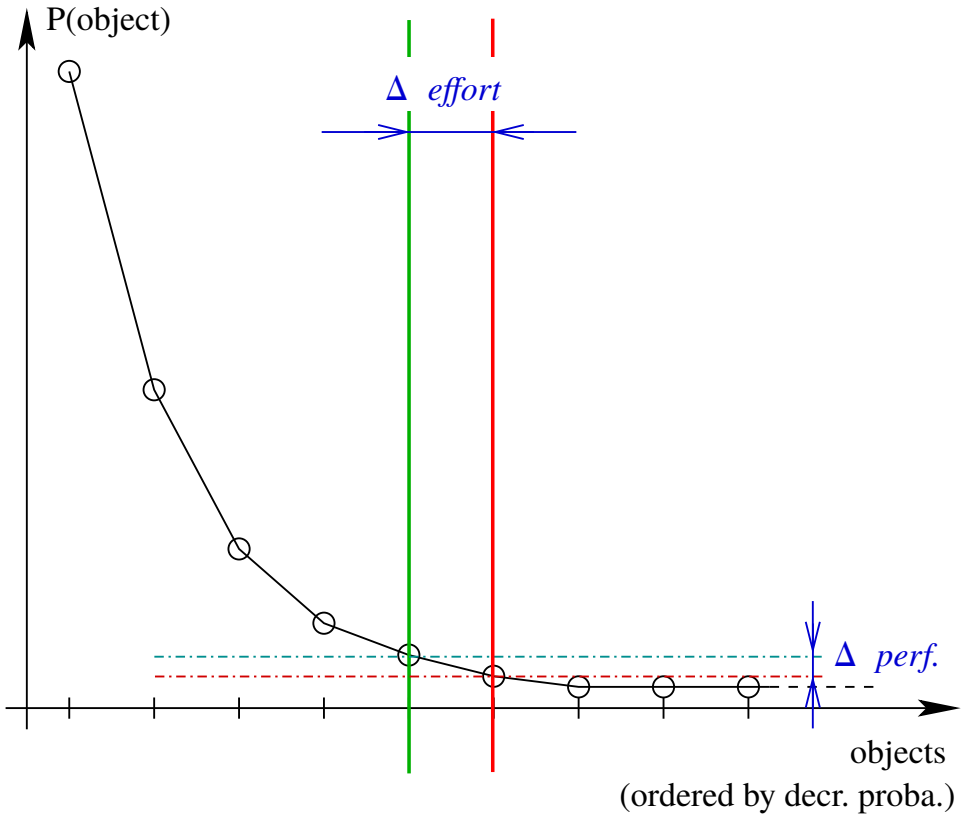
properly treat most of the corpus is thus easy for computers

☞ everybody can rapidely and easily get a “not too bad” system

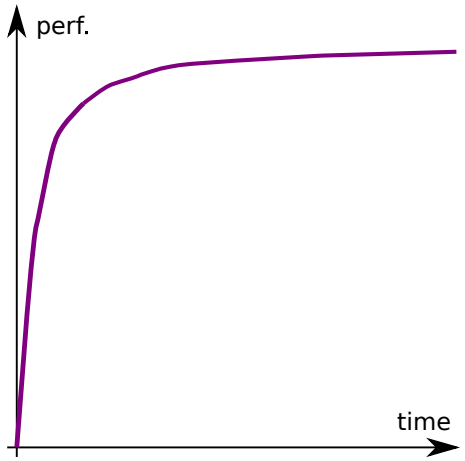


☞ The *illusion* of NLP success

The impact of power laws
(e.g. Zipf Law, “Zeta distribution”)



but getting a 0.1% improvement w.r.t actual state of the art is really not that easy!



☞ need to model the “special cases” (rare occurences)

☞ a good system **needs BOTH**! (efficient engineering/machine learning, and good NL coverage)

Need good and representative (NL) data

"*There is no data like more data*" [Mercer 85]

"*More data is more important than better algorithms*" [E. Brill]

"*We see that even out to a billion words the learners continue to benefit from additional training data*." [Banko & Brill 01]

Major issue: produce large good and representative NL resources

☞ put relevant linguistic knowledge into the learning data

☞ In a NLP system, resources production/aquisition may be at least as costly as the design of the core system itself

Why is NLP difficult?

- lack of linguistic competence
- power laws (at all levels)
- curse of dimensionality (high dimension + sparseness)
- subjectivity (Inter-Annotator (dis)Agreement)
- **multi-scale**: many levels (see next lecture), ambiguity

Keypoints

- NLP very demanded in numerous applications
- Caracteristics (conciseness and ambiguity) and functions (communication and representation) of natural language
- Trade-off between expressive power and processing time
- Linguistic resources are very important
- Corpus-based linguistics doesn't try to explain the natural language, but to improve the performances of the applications

References

[1] D. Jurafsky & J. H. Martin, *Speech and Language Processing*, Prentice Hall, 2008 (2nd edition).

[2] C. D. Manning & H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 2000.

[3] N. Indurkhya & F. J. Damerau *Handbook of Natural Language Processing*, CRC Press, 2010 (2nd edition).

[4] M. Rajman editor, "Speech and Language Engineering", EPFL Press, 2006.

[5] *Ingénierie des langues*, sous la direction de J.-M. Pierrel, Hermes, 2000.

Introduction to Natural Language Processing

PROCESSING LEVELS IN NLP

Martin Rajman

Martin.Rajman@epfl.ch

and

Jean-Cédric Chappelier

Jean-Cedric.Chappelier@epfl.ch

Artificial Intelligence Laboratory

Objectives of this lecture

- Give **general overview** of natural language processing
- Present the **main components** of an NLP system and their **relations**

Content

- Linguistic Processing Levels
- Example of an NLP architecture
- Interdependencies between processing levels

Linguistic Processing Levels

For any complete linguistic analysis, an NLP system must be able to:

- **recognize** ”words” (morpho-lexical level)

  M. O’Connel payed $ 12,000, (V.T.A. not included) with his credit card.

- **structure** the word sequences (syntactic level)

  Time flies like an arrow.

- **understand** the meaning of word sequences (semantic level)

  She ate fish with her friends / its bones.

- **contextualize** the litteral meaning (pragmatic level)

  He asked the custom officers about the taxes and payed them.

Lexical Level

Why such a **level**?

☛ To *recognize*: What is a word/token?

✏ Non-alphabetical Languages (Chinese), Languages without separators (Thaï)

✏ Ambiguous separators

- *credit card*, *due to*
- *U.N.O.*, *34,2 degrees*

✏ out-of-vocabulary forms: *dorr*, *tatcherism*, *.bat files*, *Sun*

Domain of **morphology** (study of the structure of the words) and of **lexicography** (inventory and classification of accepted forms in a language)

paradigmatic dimension of the language (.vs. **syntagmatic**)

☞ associated linguistic resources: **electronic lexica**

Syntactic level

Syntax: study of the **contraints** to be verified for a word sequence to be considered as (syntactically) ”correct” in a given language.

These constraints can be either **selectional** (agreements) or **positional**.

☞ Associated linguistic resources: (formal) **grammars**

What is Syntax useful for?

1. to solve (or reduce) some ambiguities in the lower levels:

   phonetic level: $[i][l][u][k]$ → I look
   → eye look
   → Hi! Luke
   → ...

   lexical level: he *wend* away
   → went
   wind
   end
   ...

2. to help the extraction/description/use of semantic/pragmatic facts

   Example: selectional contraints associated with the verb "*to eat*"
   ➥ animated subject
   ➥ edible object

Semantic and pragmatic levels

Semantic: meaning **out of** any context (i.e. litteral meaning)

notions of **meaning space** and of **knowledge representation**

numerical repr.	formal repr.
distance	symbolic operators

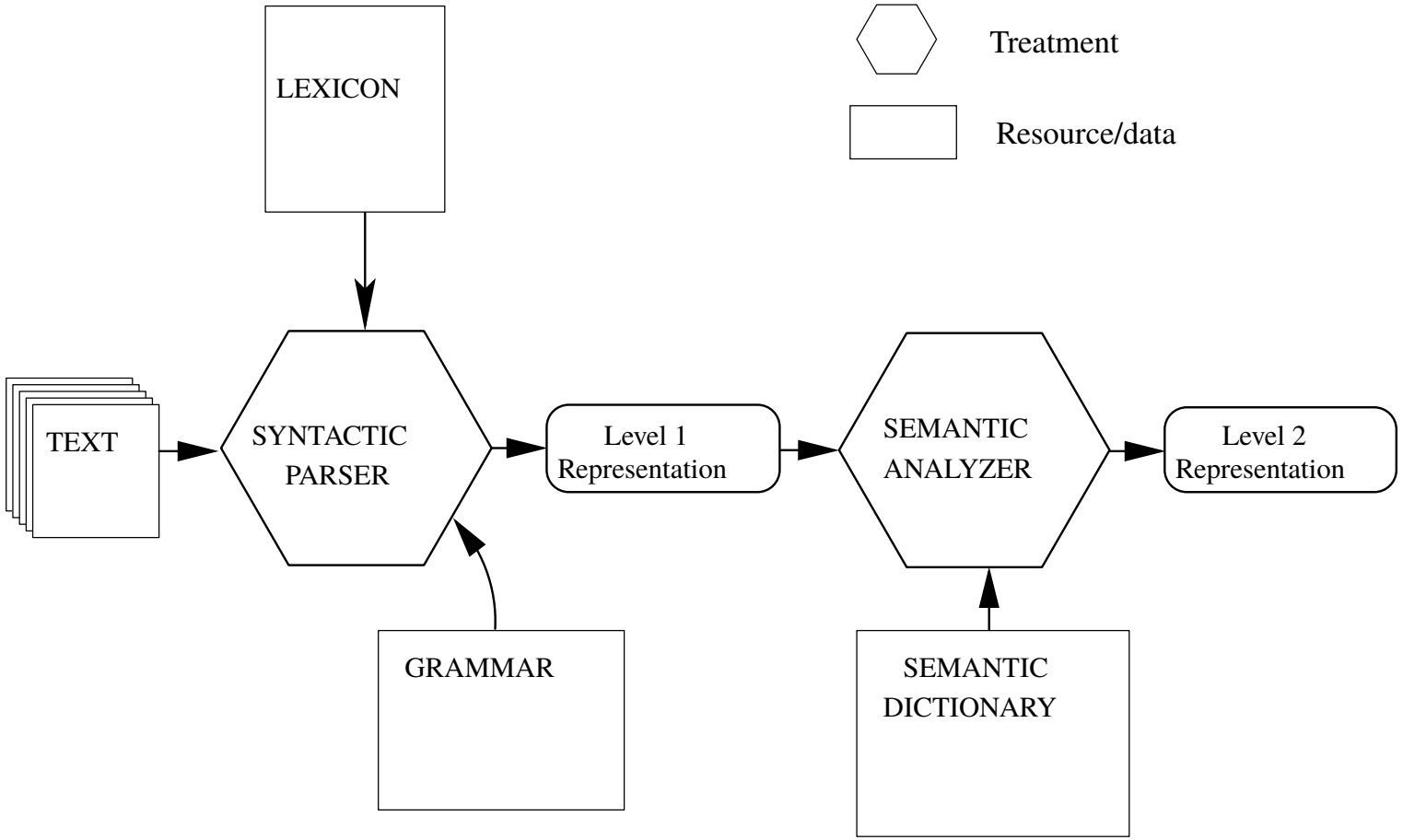
Pragmatic: meaning **within** the elocution context

Use of **knowledge representation formalisms** for formal knowledge/common sense models

Content

- Linguistic Processing Levels
- **Example of NLP architecture**
- Interdependencies between processing levels

Example of a standard NLP Architecture



Treatment .vs. Resources

Good Quality Linguistic Resources are difficult/costly to obtain/handle

☞ at least as costly as the treatments themselves

Example of resources:

- at the morpho-lexical level: **morphological rules** (grammar of the word) and **electronic lexica**
- at the syntactic level: formal **grammars** of the language
- at the semantic and pragmatic levels : **formal models of knowledge** (logical propositions, semantic networks, conceptual graphs, ...)

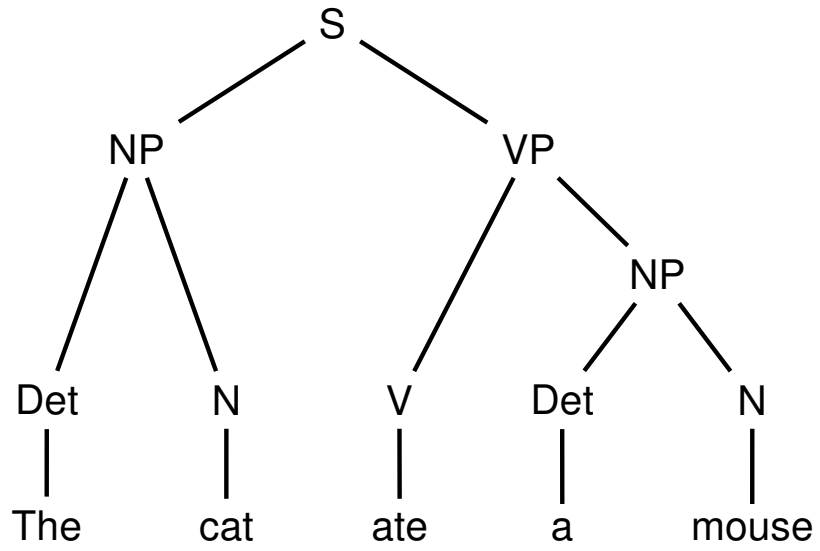
Example of a lexicon (in French)

avocat	Ncms	avocat	a v o k a
avocate	Ncfs	avocat	a v o k a t
avocates	Ncfp	avocat	a v o k a t
avocats	Ncmp	avocat	a v o k a
avoir	VaI	avoir	a v w a r
avoir	Ncms	avoir	a v w a r
avoirs	Ncmp	avoir	a v w a r
avoisina	V1is3	avoisiner	a v w a z i n a
avoisinai	V1is1	avoisiner	a v w a z i n e
avoisinaient	V1ii6	avoisiner	a v w a z i n ai
avoisinais	V1ii1	avoisiner	a v w a z i n ai
avoisinais	V1ii2	avoisiner	a v w a z i n ai
avoisinait	V1ii3	avoisiner	a v w a z i n ai
avoisinant	V1pp	avoisiner	a v w a z i n an
avoisinant	Ams	avoisiner	a v w a z i n an
avoisinante	Afs	avoisiner	a v w a z i n an t
avoisinantes	Afp	avoisiner	a v w a z i n an t
avoisinants	Ams	avoisiner	a v w a z i n an
avoisinas	V1is2	avoisiner	a v w a z i n a
avoisinasse	V1ss1	avoisiner	a v w a z i n a s
avoisinassent	V1ss6	avoisiner	a v w a z i n a s
avoisinasses	V1ss2	avoisiner	a v w a z i n a s
avoisinassiez	V1ss5	avoisiner	a v w a z i n a s y e
avoisinassions	V1ss4	avoisiner	a v w a z i n a s y on
avoisine	V1ip1	avoisiner	a v w a z i n

Example of a grammar

```
P -> GN GV {     = GN.nombre, GV.nombre,
                 < P.mode, GV.mode, }

GN -> Det N+ {   = Det.genre,  N+.genre,
                 = Det.nombre, N+.nombre,
                 < GN.genre,   N+.genre,
                 < GN.nombre,  N+.nombre, }

N+ -> ADJ N+ { * }
N+ -> N      { * }
N+ -> N ADJ  { * }

GV -> GV GN {    < GV.nombre, GV.nombre,
                 < GV.temps,  GV.temps,
                 < GV.mode,   GV.mode, }
GV -> V {        < GV.nombre, V.nombre,
                 < GV.temps,  V.temps,
                 ^ GV.mode,   pos, }
GV -> NEGpre V NEGpost {
                 < GV.nombre, V.nombre,
                 < GV.temps,  V.temps,
                 ^ GV.mode,   neg, }
```

Examples of level 1 representations



((The cat) (ate (the mouse)))

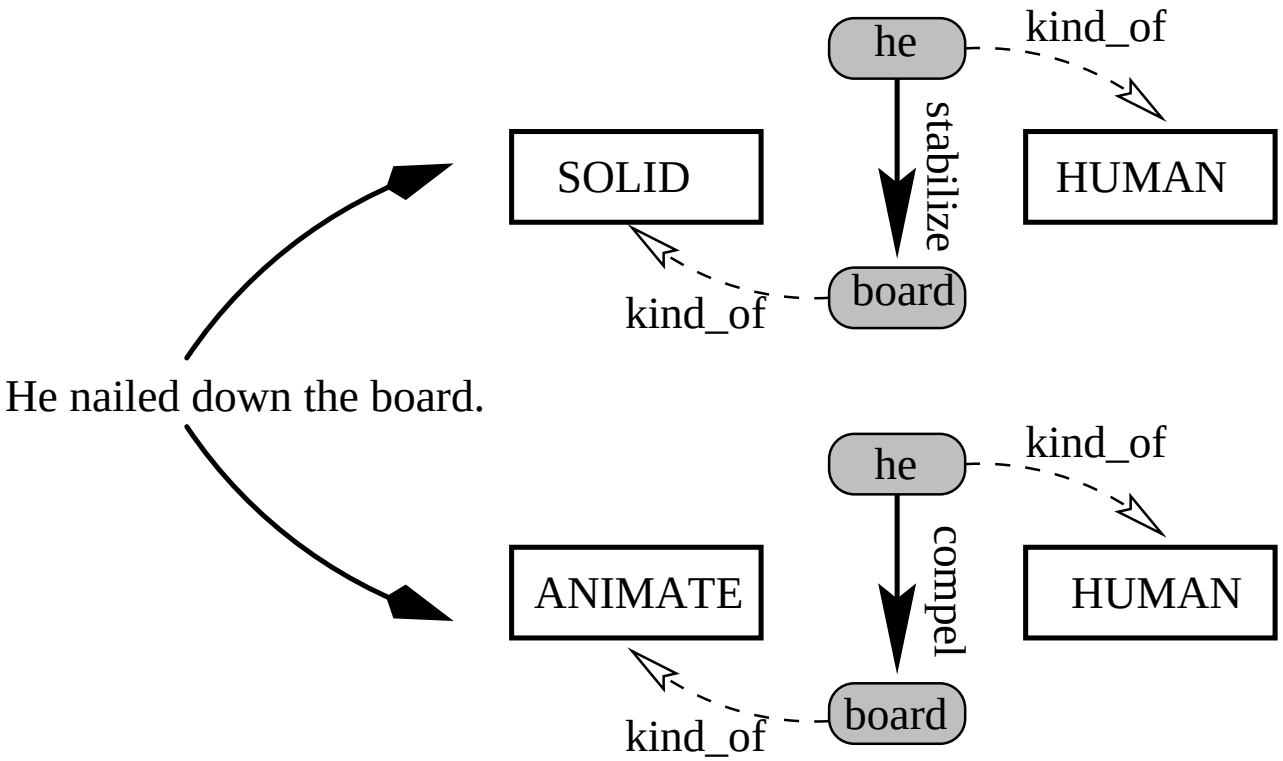
Example of semantic information (dictionary)

```
board (noun)

1 : the side of a ship

2 a : a piece of sawed lumber of little thickness and a length greatly
      exceeding its width
  b plural : STAGE

3 a archaic : TABLE
  b : a table spread with a meal
  c : daily meals especially when furnished for pay
  d : a table at which a council or magistrates sit
  e (1) : a group of persons having managerial, supervisory,
          investigatory, or advisory powers
          <board of directors> <board of examiners>
...

5 a : a flat usually rectangular piece of material (as wood) designed for a
      special purpose: as SPRINGBOARD, SURFBOARD
...
```

Example of semantic knowledge

```
board:N -> league -> humans -> animate
board:N -> solid, object    -> inanimate

nail_down(X,Y) :- stabilize(X,Y), human(X), solid(Y).
nail_down(X,Y) :- compel(X,Y),    human(X), animate(Y).
```

Example of level 2 representation



Content

- Linguistic Processing Levels
- Example of NLP architecture
- **Interdependencies between processing levels**

Interdependencies between processing levels

Interdependencies between the lexical level and the other levels:

Example of spelling error correction:

*the **tost** of the coin*

tost $\rightarrow$ lost : syntax

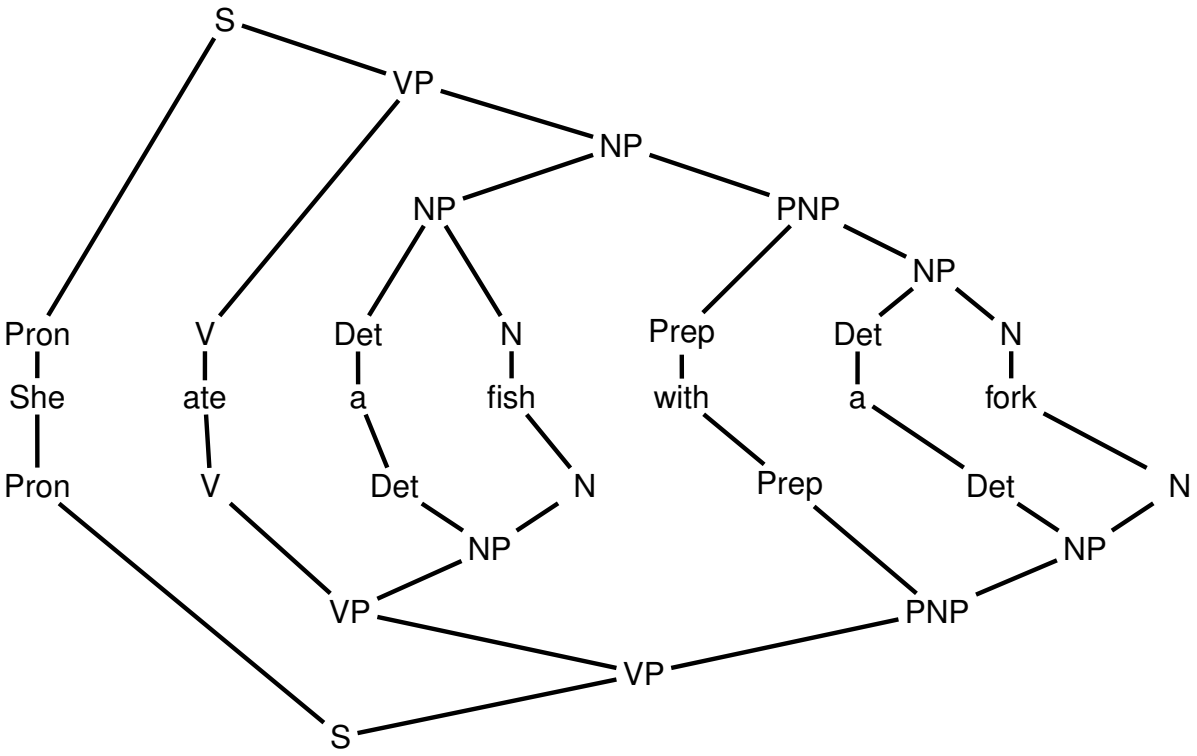
$\rightarrow$ toast : semantics

$\rightarrow$ cost : pragmatics

$\rightarrow$ toss : pragmatics

syntactic-semantics

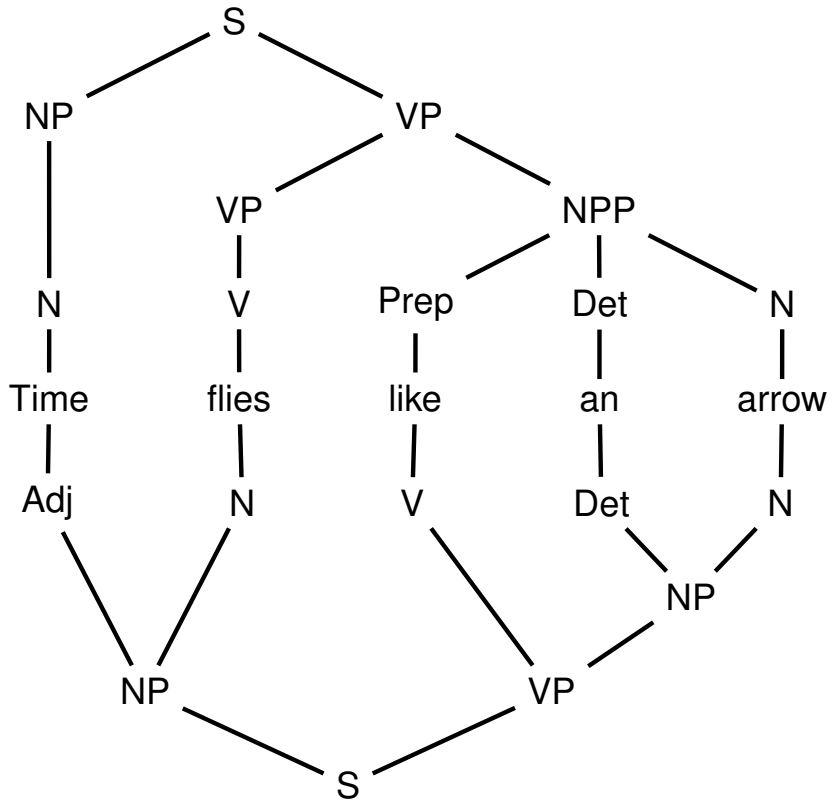
She ate a fish with a fork.



semantic knowledge...

syntactic-pragmatics dependency

...but with: *Time flies like an arrow.*



pragmatic knowledge

Keypoints

- Main stages of linguistic analysis and architecture of an NLP system
- Components of an NLP system (word recognition and structuring, phrase understanding and contextualization) and their implementation
- Interdependence between NLP components (recognition conditionned by structuring, structuring guided by the meaning and the context)

References

[1] D. Jurafsky & J. H. Martin, *Speech and Language Processing*, Prentice Hall, 2008 (2nd edition).

[2] C. D. Manning & H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 2000.

[3] N. Indurkhya & F. J. Damerau *Handbook of Natural Language Processing*, CRC Press, 2010 (2nd edition).

[4] M. Rajman editor, "Speech and Language Engineering", EPFL Press, 2006.

[5] *Ingénierie des langues*, sous la direction de J.-M. Pierrel, Hermes, 2000.