

# Natural Language Processing: Introduction

Roberto Navigli

DIPARTIMENTO  
DI INFORMATICA



SAPIENZA  
UNIVERSITÀ DI ROMA

## Your Instructor

- Associate Professor in the Department of Computer Science (Sapienza)
- Home page: <http://wwwusers.di.uniroma1.it/~navigli>
- Email: [navigli@di.uniroma1.it](mailto:navigli@di.uniroma1.it)



## What is Natural Language Processing (NLP)?

- The branch of information science that deals with natural language information [WordNet]



## But... what is Information Science?

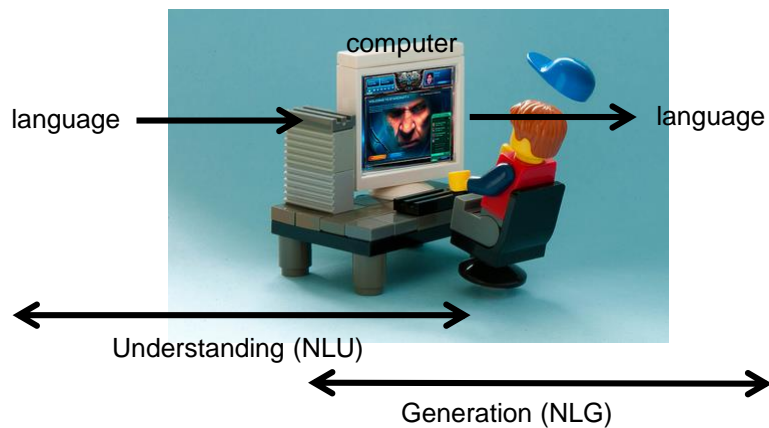
- Information science is an interdisciplinary science primarily concerned with the **analysis, collection, classification, manipulation, storage, retrieval and dissemination** of information [Wikipedia]

## NLP: an Interdisciplinary Area

- Artificial Intelligence
- Computer Science
- Linguistics
- Psychology
- Logic
- Statistics
- Cognitive science
- Neurobiology
- ...

## What is Natural Language Processing II

- The use of natural language by computers as **input** and/or **output**



## Natural Language Processing and Artificial Intelligence

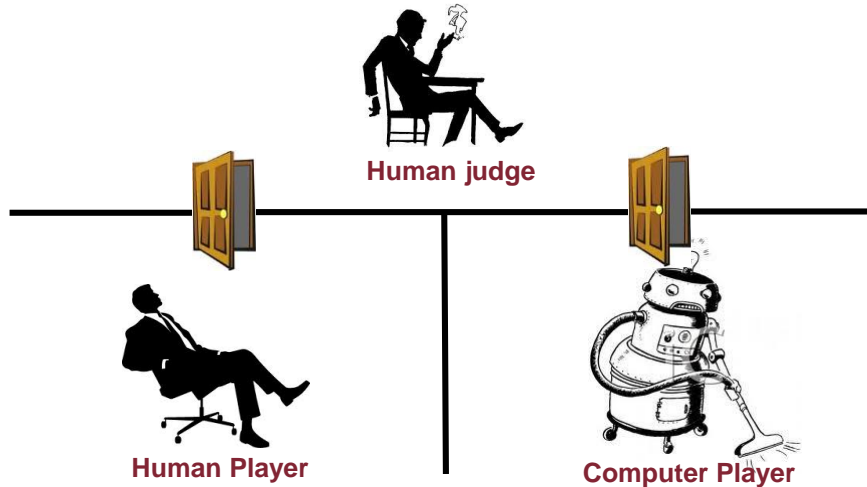
- NLP is a branch of Artificial Intelligence (AI)
- Better: NLP is the branch of AI **dealing with human language**
- **Intelligence** comprises capacities for:
  - Abstract thought
  - Understanding
  - Communication
  - Reasoning
  - Learning
  - Planning
  - Emotions
  - Problem solving
- How do we know whether a living being/system is **intelligent**?
- **Idea**: use language to test!

## Turing Test (1950)

- A test of a **machine's** ability to **demonstrate intelligence**
- Introduced in 1950 by **Alan Turing**
- "I propose to consider the question, 'Can machines think?'" Since "thinking" is difficult to define, Turing chooses to "replace the question by another, which is closely related to it and is expressed in relatively unambiguous words. [...] Are there imaginable digital computers which would do well in the imitation game?"
  - Alan Turing, "Computing Machinery and Intelligence" (1950)
- Inspired by a **party game**, known as the "imitation game" (a man vs. a woman)

## Turing Test (1950)

- A test of a machine's ability to demonstrate intelligence



## Turing Test (1950)

- A human judge engages in a (written) natural language conversation with one human and one machine
- The players try to appear human
- All participants are separated from each another
- The judge tries to determine which player is a computer and which is a human
- Assumption: NLP is AI-complete!
- In other words, if we solve NLP, we are able to solve AI

## ELIZA (1966)

- An early example of computer program performing primitive natural language processing
  - Written at MIT by Joseph Weizenbaum (1966)
- Processes human replies to questions
- Uses simple parsing and substitutes keywords into template phrases

```
HI! I'M ELIZA. WHAT'S YOUR PROBLEM?  
? IM ADDICTED TO MOBYGAMES  
DID YOU COME TO ME BECAUSE YOU ARE ADDIC  
TED TO MOBYGAMES  
? YES  
ARE YOU SURE?  
? NOT ANYMORE  
ARE YOU SAYING NO JUST TO BE NEGATIVE?  
? NO  
YOU ARE BEING A BIT NEGATIVE.  
? YOU ARE BEING ANNOYING  
WHAT MAKES YOU THINK I AM BEING ANNOYING  
? YOU DONT CONTRIBUTE ANYTHING USEFUL TO  
THE CONVERSATION  
WE WERE DISCUSSING YOU--NOT ME.  
? THATS NOT A QUESTION  
WHY NOT?  
? BECAUSE ITS A STATEMENT  
IS THAT THE REAL REASON?  
? THERE ALSO ISNT A QUESTION MARK  
SAY DO YOU HAVE ANY PSYCHOLOGICAL PROBLE  
MS?  
? I FEEL ANNOYED WHEN TALKING TO A POOR  
AI
```

## Loebner Prize Gold Medal

- \$100,000 and a **Gold Medal** for the first computer whose responses were indistinguishable from a human's
- <http://www.loebner.net/Prizef/loebner-prize.html>



## The Chinese Room (1980)

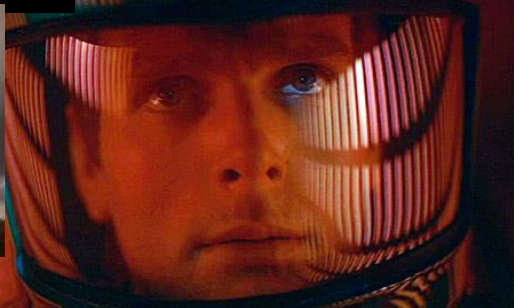
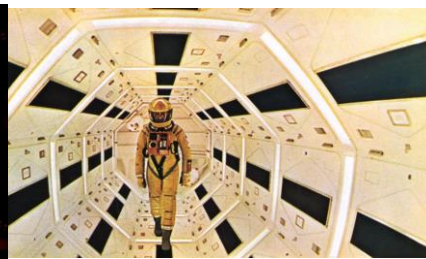
- **John Searle** argued against the Turing Test in his 1980 paper "Minds, Brains and Programs"
- **Programs** (e.g., ELIZA) could pass the Turing Test simply by **manipulating symbols** they do not understand
- Assume you act as a computer by **manually executing a program** that **simulates** the behavior of a native Chinese speaker

## The Chinese Room (1980)

- Assume you are in a **closed room** with a book containing the computer program
- You receive **Chinese characters** through a slot in the door and process them according to the program's instructions and produce Chinese characters as output
- Would this mean that you **understand**?
- Would it mean you can **speak Chinese**?
- "I can have any formal program you like, but I still understand nothing."



## The science fiction dream!

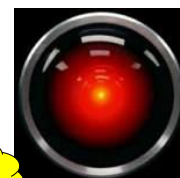






## What knowledge does HAL 9000 need?

- HAL (Heuristically programmed ALgorithmic computer)



**Dave Bowman:** Hello, HAL. Do you read me, HAL?

**HAL:** Affirmative, Dave. I read you.

**Dave Bowman:** Open the pod bay doors, HAL.

**HAL:** I'm sorry, Dave. I'm afraid I can't do that.

**Dave Bowman:** What's the problem?

**HAL:** I think you know what the problem is just as well as I do.

**Dave Bowman:** What are you talking about, HAL?

**HAL:** This mission is too important for me to allow you to jeopardize it.

**Dave Bowman:** I don't know what you're talking about, HAL.

syntax

phonetics & phonology

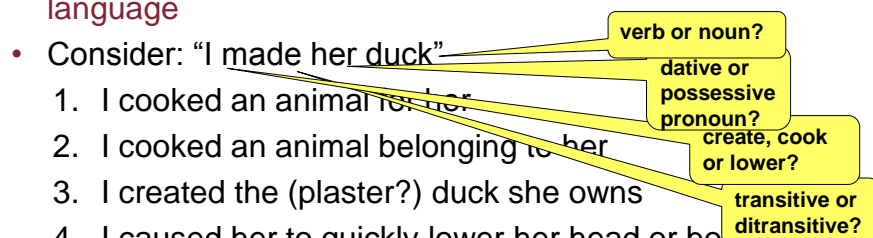
semantics

discourse

pragmatics

morphology

## Why is NLP so hard?

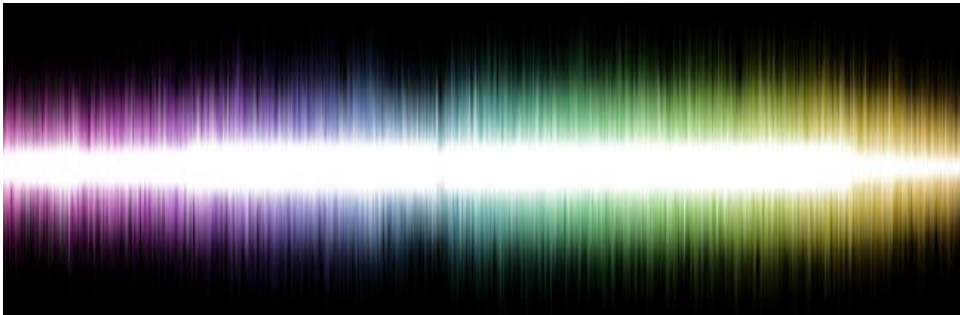
- The answer is: **ambiguity** at the different **levels of language**
  - Consider: “I made her duck”
    1. I cooked an animal for her
    2. I cooked an animal belonging to her
    3. I created the (plaster?) duck she owns
    4. I caused her to quickly lower her head or body
    5. I magically transformed her into a duck [ditransitive]
  - Further ambiguity of spoken language:  
“eye made her duck” ...
- 

## The aim of NLP

- Resolving such ambiguities by means of **computational models and algorithms**
- For instance:
  - **part-of-speech tagging** resolves the ambiguity between duck as *verb* and *noun*
  - **word sense disambiguation** decides whether *make* means *create* or *cook*
  - **probabilistic parsing** decides whether *her* and *duck* are part of the same syntactic entity

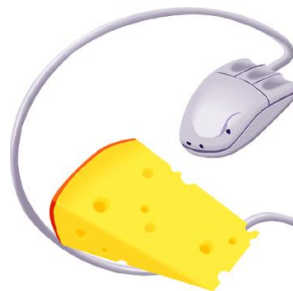
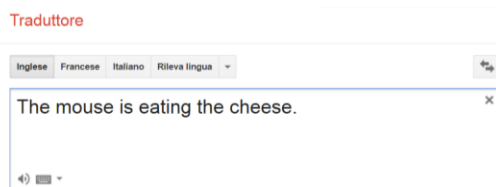
## Why Aren't We All Talking *With* Our Devices Yet?

- “Because it takes **more than just understanding a bunch of words** to provide a good voice user interface — especially on a mobile phone. We have to understand intent. But there are other factors at play here besides the technology of speech: output, interaction, and context.” [Chris Schmandt, MIT Media Lab]



## Examples of the importance of NLP

### 1. Machine Translation



## Computer-Assisted Translation (CAT)



## Examples of the importance of NLP

### 2. Text Summarization

Follow these simple steps to create a summary of your text.

**Step 1**  
Type or paste your text into the box.

Arm wrestling is a type of wrestling (a combat sport) with two participants. Each participant places one arm (either the right or left, but both must be the same) on a surface with their elbows bent and touching the surface, and they grip each other's hand. The goal is to pin the other's arm onto the surface, with the winner's arm over the loser's arm.

Description

Various factors can play a part in one's success in arm wrestling. Technique and overall arm strength are the two greatest contributing factors to winning an arm wrestling match. Other factors such as the length of an arm wrestler's arm, his/her muscle and arm mass/density, hand grip size, wrist endurance and flexibility, reaction time, as well as countless other traits, can add to the advantages of one arm wrestler over another.[citation needed]. It is sometimes used to prove who is stronger between two or more people. In competitive arm wrestling, as sanctioned by the United States Armwrestling Federation (USAF), arm wrestling is performed with both competitors standing up with their arms placed on a tournament arm wrestling table.[citation needed]. Arm wrestling tournaments are also divided into weight classes as well

**Step 2**  
Drag the slider, or enter a number in the box, to set the percentage of text to keep in the summary.

4 %

**Step 3**  
Read your summarized text. If you would like a different summary, repeat Step 2. When you are happy with the summary, copy and paste the text into a word processor, or [text to speech program](#), or [language translation tool](#).


The second generic system or style of Arm wrestling is known as outside arm wrestling "the top roll" or "top rolling", while the "tricep press", "shoulder pressing", or "shoulder rolling" is often described as the third generic system or style of arm wrestling.[citation needed] and certain arm wrestlers depend on the straps[clarification needed] such as Jason Vale who won the 1997 Petaluma World Championships in the super heavy weight class at only 175 pounds using the strap technique.[citation needed]  
The contestant on the right is in an injury-prone or "break arm" position.

## Examples of the importance of NLP

### 3. Personal assistance

Google Now. Le informazioni giuste al momento giusto.

Schede utili con le informazioni che ti servono nel corso della giornata, visualizzate ancor prima che tu le chiedi.



L'app Google

Download on the App Store

Hi, I'm Cortana.

Natural Language Processing: An Introduction  
Roberto Navigli

## Examples of the importance of NLP

### 4. Information Extraction / Machine Reading



#### Open Information Extraction



**Example Queries:**

- What kills bacteria?
- Who built the Pyramids?
- What did Thomas Edison invent?
- What contains antioxidants?

**Typed Example Queries:**

- What countries are located in Africa?
- What actors starred in which films?
- What is the symbol of which country?
- What foods are grown in which countries?
- What drug ingredients has the FDA approved?

Argument 1:  
what/who

**NELL: Never-Ending Language Learning**

Can computers learn to read? We think so. "Read the Web" is a research project that attempts to create a computer system that learns over time to read the web. Since January 2010, our computer system called NELL (Never-Ending Language Learner) has been running continuously, attempting to perform two tasks each day:

- First, it attempts to "read," or extract facts from text found in hundreds of millions of web pages (e.g. `playsInstrument(George_Harrison, guitar)`).
- Second, it attempts to improve its reading competence, so that tomorrow it can extract more facts from the web, more accurately.

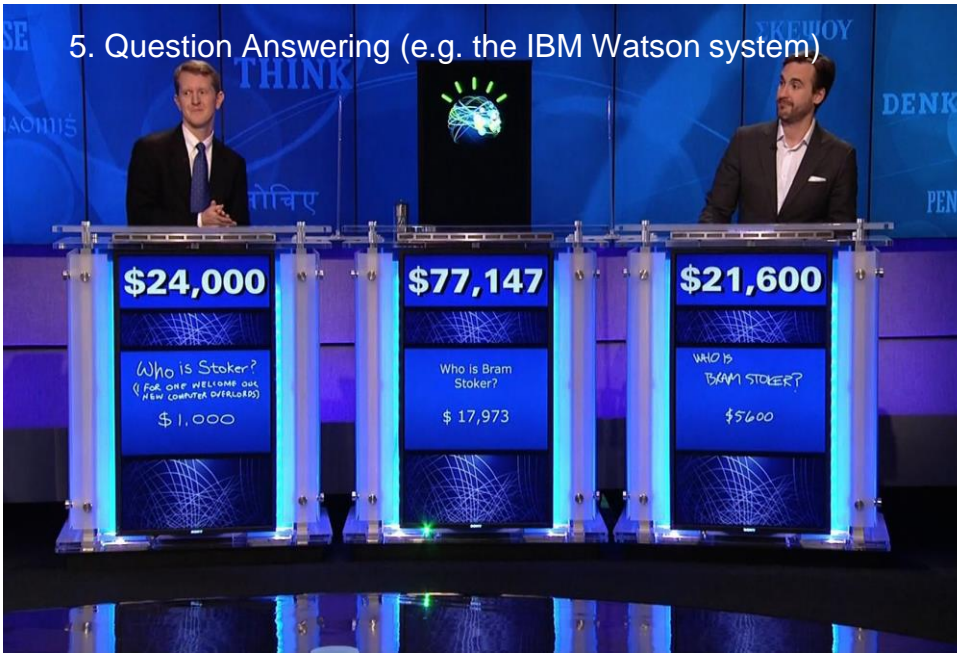
So far, NELL has accumulated over 50 million candidate beliefs by reading the web, and it is considering these at different levels of confidence. NELL has high confidence in 1,986,058 of these beliefs — these are displayed on this website. It is not perfect, but NELL is learning. You can track NELL's progress below or [@cmunell on Twitter](#), browse and download its [knowledge base](#), read more about our [technical approach](#), or join the [discussion group](#).

**Browse the Knowledge Base!**

Natural Language Processing  
Roberto Navigli

## Examples of the Importance of NLP

### 5. Question Answering (e.g. the IBM Watson system)



## Examples of the importance of NLP

### 6. Information Retrieval





Web Shopping Images News Videos More Search tools

About 1,070,000 results (0.35 seconds)

[Which painkiller? - Live Well - NHS Choices](#)

[www.nhs.uk/Livewell/Pain/.../Whichpainkiller.as...](#) National Health Service  
If you take them for long periods, there's an increased risk of **stomach upset**, including bleeding, and kidney and heart problems. **Don't** take more than the ...

[Are there any anti-inflammatory drugs that don't have stomach ...](#)

[www.arthritisresearchuk.org/.../any-nsaids-without-...](#) Arthritis Research UK  
Are there any anti-inflammatory drugs that **don't** have **stomach**-related side-effects? ... Do you know of any anti-inflammatories I can take that won't **upset** my ulcers ... In these situations we recommend **painkillers** such as the one you are taking.

[Any pain relievers that don't upset the stomach if taken often ...](#)

[www.godlikeproductions.com/forum1/message1121319/pg1](#)  
Jul 3, 2010 - 31 posts - 9 authors  
Aspirin, advil, aleve, tylenol all mess my **stomach** up over time. Any safe pain pills out there prescription or non prescription? And please **don't** ...

[Pain Killers Comparison Chart - Painkiller Summary](#)

[www.vaughns-1-pagers.com/medicine/painkiller-comparison.htm](#)  
A summary chart of **pain killers**, ranked by effectiveness for back pain. Both OTC and prescription ... **upset stomach**, not for last trimester .... I **don't** know. But it is a ...

[How taking painkillers can destroy your stomach lining in days | Mail ...](#)

[www.dailymail.co.uk/.../How-taking-painkillers-destroy-stoma...](#) Daily Mail  
Sep 26, 2011 - Claire Calder's (pictured) stomach lining was so damaged from taking ... non-ulcer **dyspepsia** — a condition that causes chronic stomach pain. ... Doctors have yet to establish why some people react so badly and others **don't**.

[In search of painkillers that don't damage the stomach \[Archive ...](#)

[harris.straightdope.com](#) > General Questions > The Straight Dope

Web Images Shopping News Videos More Search tools

About 1,790,000 results (0.32 seconds)

Searches related to **Painkillers such as acetaminophen stomach**

- Arthritis** Painful inflammation and stiffness of the joints
- Liver disease** A broad term describing any single number of diseases a...
- Reyes syndrome** A life-threatening metabolic disorder in young children, of...
- Flu** An infectious disease caused by ma viruses of the family...
- Heart attack** A sudden and sometimes fatal occurrence of coronary th...

Drawn from at least 10 websites, including [drugs.com](#) and [wikipedia.org](#) - [How this works](#)

[Osteoarthritis Medications Options: Analgesics, NSAIDs ...](#)

[www.spine-health.com](#) > Conditions > Arthritis  
**Acetaminophen** does not reduce inflammation, but is an effective pain reliever and is less likely to cause **stomach** problems than NSAIDs (such as **ibuprofen** or ...

[Picking a pain reliever: which one should you take? - NYU Langone ...](#)

[www.med.nyu.edu/content?ChunkID...](#) NYU Langone Medical Center  
... pain relievers, such as aspirin, **acetaminophen** (Tylenol), **ibuprofen** (Advil, Motrin), ... To minimize **stomach** upset, some aspirin products are buffered with an ...

[Non-steroidal anti-inflammatory drug - Wikipedia, the free ...](#)

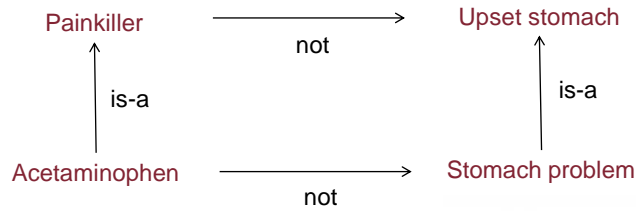
[en.wikipedia.org/wiki/Non-steroidal\\_anti-inflammatory\\_drug](#) Wikipedia  
Over the past decade, deaths associated with **gastric** bleeding have declined. ... Pain relievers such as **paracetamol** (also known as **acetaminophen**) or drugs ...

[SHOULD I TAKE TYLENOL, ADVIL OR ASPIRIN? | Science Creative ...](#)

[www.scq.ubc.ca/should-i-take-tylenol-adv...](#) University of British Columbia  
Nov 21, 2006 - However, some non-prescription **painkillers**, such as **Tylenol**, **Advil** and **Aspirin** are also ... It rarely causes **stomach** upset or allergic reactions.

[Arthritis: What can prevent stomach ulcers caused by painkillers and ...](#)

## I need to reason...



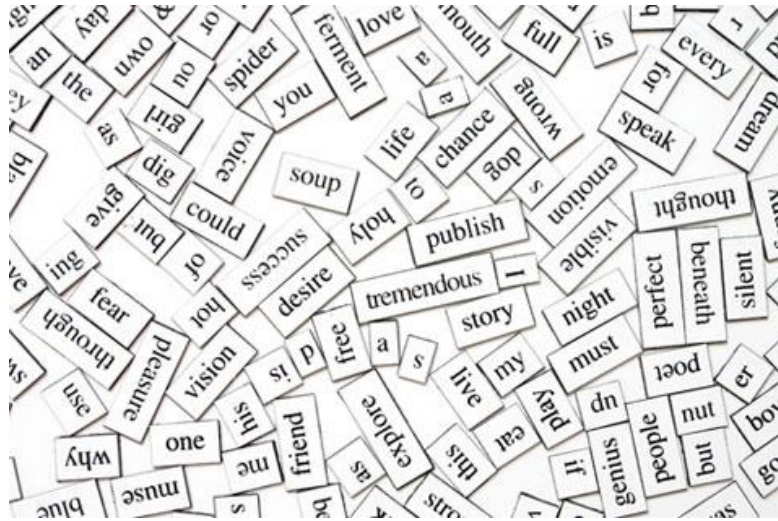
## Natural Language Processing in Brief

- Morphological Analysis
- Language modeling
- Part-of-speech tagging
- Syntactic Parsing
- Computational Lexical Semantics
- Statistical Machine Translation
- Discourse and Dialogue
- Text Summarization
- Question Answering
- Information Extraction and Text Mining
- Speech Processing



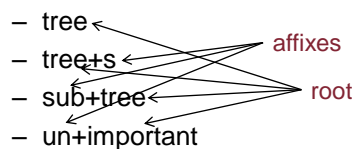


## We are talking about words!



## What are words?

- Basic **building block** of language
- Every human language, either spoken or written, is **composed of words**
- A **word** is the **smallest free form** that can be uttered in isolation with **semantic** and **pragmatic** content
- Made up of **morphemes** (smallest component of **word** that has semantic meaning)



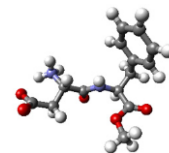
## We need to perform a morphological analysis

- “Morphology is the study of the way words are built up from smaller meaning-bearing units” (Jurafsky & Martin, 2000)
- The meaning-bearing units are called morphemes
- Two main types of morphemes:
  - Stem or root: the main morpheme of a word
  - Affixes: prefixes (re-write), suffixes (beauti-ful-ly), infixes and circumfixes
- In order to detect these components we need to perform morphological parsing

## The concept of parsing



Input



Structure for the input

## Morphological Parsing

Input	Morphologically Parsed Output
beagles	beagle +N +PL
cities	city +N +PL
buy	buy +N +SG or buy +V
buying	buy +V +PRES-PART
bought	buy +V +PAST-PART or buy +V +PAST

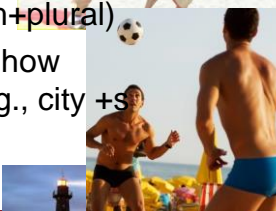
- What do we need to build a **morphological parser**?

## Ingredients for a Morphological Parser

- **Lexicon**: the list of stems and affixes

Stem	Part of speech
------	----------------

- **Morphotactics**: the model of morpheme ordering in the language of interest (e.g., main stem+plural)
- **Orthographic rules**: spelling rules about how morphemes combine to form a word (e.g., city +s -> cities)



## Ingredients for a Morphological Parser

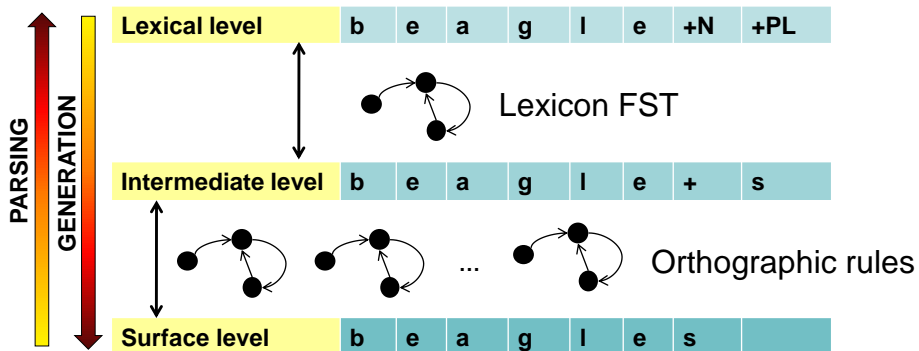
- **Lexicon:** the list of stems and affixes

Stem	Part of speech
beach	N
beachwards	ADV
beachwear	N
beachy	ADJ
beacon	N

- **Morphotactics:** the model of morpheme order in the language of interest (e.g., main stem-  
affix)
- **Orthographic rules:** spelling rules about how morphemes combine to form a word (e.g., city  
-> cities)



## Three levels: lexical, intermediate and surface level



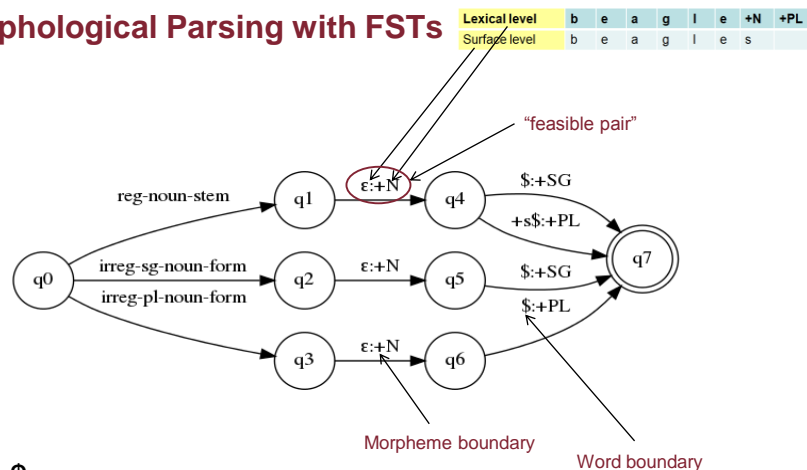
## Morphological Parsing with Finite State Transducers

- We would like to keep **distinct** the **surface** and the **lexical** levels
- We need to build mapping rules between **concatenation of letters** and **morpheme+feature** sequences

Lexical level	b	e	a	g	l	e	+N	+PL
Surface level	b	e	a	g	l	e	s	

- A **finite-state transducer** implements **two-level morphology** and maps between one set of symbols to another
  - Done using a (two-tape) **finite-state automaton**
  - Recognizes or generates **pairs** of strings

## Morphological Parsing with FSTs



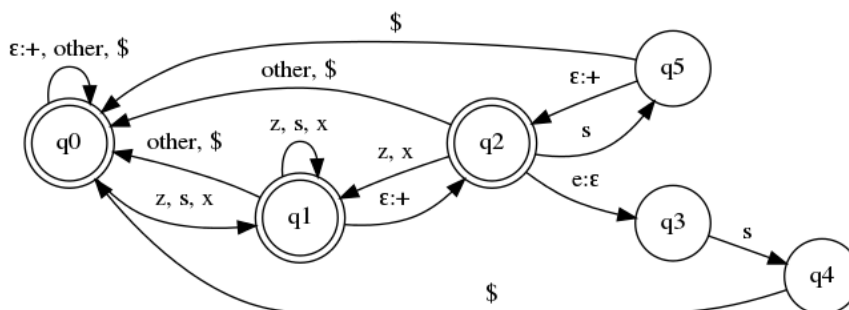
- **+s\$** means:
  - “+” (morpheme boundary), “s” (the morpheme), “\$” word boundary

## But: Just concatenating morphemes doesn't always work!

- **box+s = boxs**, rather than **boxes**
- **boxes** wouldn't be recognized!
- Why? Because of a **spelling change at morpheme boundaries**
- We need to introduce **spelling (or orthographical) rules**
  - And **implement these rules** as FSTs

Rule	Description	Example
Consonant doubling	1 consonant doubled before –ing or –ed	beg/begging, embed/embedded
E deletion	e taken out before –ing or -ed	make/making
E insertion	e added after –s, -z, -x, -ch, -sh before s	watch/watches
Y replacement	-y replaced by –ies before –s, -i before -ed	try/tries, try/tried, city/cities

## Example: transducer for the "E insertion" rule



- So one can build a transducer for each **spelling and orthographical** rule
- For example: foxes -> fox+s  
(q0 -> q0 -> q1 -> q2 -> q3 -> q4 -> q0)

## Where do we go from here?

- We are now able to process text at the morphological level
- We can work on **word combinations**
- For instance, from The Telegraph:
  - Escort claims Berlusconi's 'bunga bunga' parties full of young...
- **What comes next?**
  - Old women? Boys? Girls?
  - **It depends! On what?**

## 1-grams (unigrams): just a single word

- Absolute count of each word (e.g. on the Web):

</S>	95119665584
<S>	95119665584
,	30578667846
.	22077031422
<UNK>	21594821357
the	19401194714
-	16337125274
of	12765289150
and	12522922536
:	12255665115
to	11557321584
)	9036544694
(	8912668768
a	7841087012
in	7490628883

## 2-grams (bigrams): sequences of two words

- Absolute count of two words (e.g. on the Web):

young gipsy	267
young gir	1203
young gir.s	79
young giraffe	817
young giraffes	288
young giral	77
young girel	245
young girels	227
young girls	266
young girls	59
young girls	379
young girl	1716008
young girl'	64
young girl.I	138
young girla	117

## 3-grams (trigrams): sequences of 3 words



of young bone	75
of young bones	56
of young boobs	66
of young book	177
of young born	362
of young botanists	41
of young bovine	41
of young bowlers	147
of young boxers	184
of young boy	11928
of young boys	40490
of young brain	60
of young brains	220
of young branches	421
of young branchlets	65



of young german	80
of young giant	245
of young giants	55
of young gibbons	48
of young gifted	589
of young ginger	96
of young girl	11631
of young girlfriends	153
of young girlhood	48
of young girls	86186
of young girlson	101
of young global	119
of young globular	166





## Word prediction and N-gram models

- We can create **language models**, called **N-gram models**
  - they **predict the next word** from the **previous N-1 words**
  - they define **probability distributions** over strings of text
- Useful in:
  - Speech recognition
  - Handwriting recognition
  - Machine translation
  - Spelling correction
  - Part-of-speech tagging

## Simple N-grams

- Our aim is to **compute the probability** of a word given some history:

$$P(w|h)$$

- For instance:
  - $P(\text{rapa}|\text{qui non l'ha capito nessuno che questa è una}) =$   
 $C(\text{qui non l'ha capito nessuno che questa è una rapa})/$   
 $C(\text{qui non l'ha capito nessuno che questa è una})$
- **How easily** can we calculate this probability?

## 1) It depends on the corpus

- With a **large corpus**, such as the Web, we can compute these counts
- But: try yourself!



- Bad luck?
- The Web is **not big enough (!)** to provide good estimates for most counts

## 2) Language is infinite

- You are a good student
  - About 4,630,000 results (0.19 seconds)
- You are a **very** good student
  - About 2,040,000 results (0.36 seconds)
- You are a **very very** good student
  - 7 results (0.26 seconds)
- You are a **very very very** good student
  - 1 result (0.25 seconds)
- You are a **very very very very** good student
  - 0 results!



Too good  
for the  
Web!

## So what is a language model?

A language model is a probability distribution over word sequences

- P("You are a good student") will be **high**
- P("You are a very very very very good student") will be **very low**

## We need to estimate probabilities

- Chain rule of probability:

$$\begin{aligned} P(w_1 \dots w_n) &= P(w_1)P(w_2 | w_1)P(w_3 | w_1^2) \dots P(w_n | w_1^{n-1}) \\ &= \prod_{k=1}^n P(w_k | w_1^{k-1}) \end{aligned}$$

- Not enough - we need to **approximate**:

$$P(w_n | w_1^{n-1}) \approx P(w_n | w_{n-1})$$

- Independence assumption: **Markov assumption** of order **N-1**
- How to **estimate** these bigram probabilities (N=2)?

## Calculating the Relative Frequency Estimate

- Bracket sentences with <s> and </s>
- Count the frequency of each bigram
- We estimate the **bigram probabilities** by normalizing counts from a corpus:

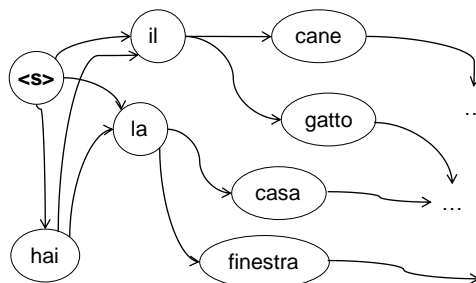
$$P(w_n | w_{n-1}) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1}w)} = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

- General case with **N-gram probabilities**:

$$P(w_n | w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}w_n)}{C(w_{n-N+1}^{n-1})}$$

## Bigram Models are Markov chains

	<s>	il	...	cane
<s>	0	0.3	...	0.01
il	0	0	...	0.2
...	0	...	...	...
cane	0	0.1	...	0.01



A random process usually characterized as memoryless: the next state depends only on the current state

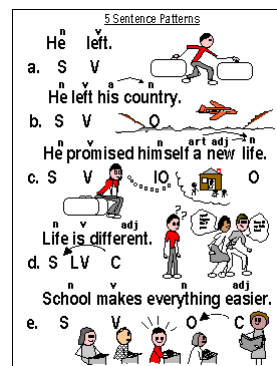
## Word classes / Parts Of Speech (POS) / Lexical Tags

- Why do we need **word classes**?
- They give **important information** about the **word** and its **neighbours**
- He is running the **race**      He decided to **race** for the job



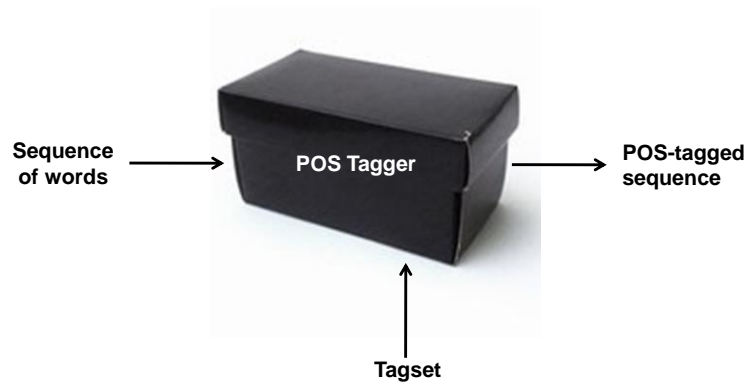
## Word classes / Parts Of Speech (POS) / Lexical Tags

- Why do we need **word classes**?
- They give **important information** about the **word** and its **neighbours**
- Useful for **recognizing speech** and **correcting spelling errors**:
  - What **is likely** to come after an adjective?
  - a verb? a preposition? a noun?



## Part-of-Speech Tagging

- It consists of assigning a part-of-speech tag to each word in a text automatically



## Part-of-Speech Ambiguity

- Unfortunately, the same word can be tagged with different POS tags:
  - How to increase the water pressure from a **well**?
  - Tears **well** in her eyes
  - The wound is nearly **well**
  - The party went **well**
- Part-of-Speech tagging is a **disambiguation task!**



## An Example: an English sentence

sentence: The oboist Heinz Holliger has taken a hard line about the problems .  
original: DT NN NNP NNP VBZ VBN DT JJ NN IN DT NNS .  
universal: DET NOUN NOUN NOUN VERB VERB DET ADJ NOUN ADP DET NOUN .

Example English sentence with its language specific and corresponding universal POS tags.

## Stochastic Part-of-Speech Tagging (since 1970s)

- Stochastic POS tagging uses probabilities to tag
- **Idea:** use **Hidden Markov Models** to select the most-likely tags for a given sequence of words

$$\hat{t}_1^n = \arg \max_{t_1^n \in \text{Tagset}^n} P(t_1^n | w_1^n)$$

- But how can we **calculate** these probabilities?

## Holy Bayes!

- Remember?

$$P(x | y) = \frac{P(y | x)P(x)}{P(y)}$$



- Let's apply **Bayes' Theorem** to our formula:

$$\hat{t}_1^n = \arg \max_{t_1^n \in \text{Tagset}^n} \frac{P(w_1^n | t_1^n)P(t_1^n)}{P(w_1^n)} = \arg \max_{t_1^n \in \text{Tagset}^n} \underbrace{P(w_1^n | t_1^n)}_{\text{likelihood}} \underbrace{P(t_1^n)}_{\text{prior}}$$

- Still **hard** to compute!

## HMM taggers make two simplifying assumptions

- 1) The probability of a word depends only on its own part-of-speech tag

$$P(w_1^n | t_1^n) = P(w_1 | t_1^n)P(w_2 | w_1, t_1^n) \dots P(w_n | w_1^{n-1}, t_1^n) \approx \prod_{i=1}^n P(w_i | t_i)$$

- 2) The probability of a tag appearing depends only on the previous tag (**bigram assumption**)

$$P(t_1^n) \approx \prod_{i=1}^n P(t_i | t_{i-1})$$



## The two simplifying assumptions in action

$$\hat{t}_1^n = \arg \max_{t_1^n \in \text{Tagset}^n} P(t_1^n | w_1^n) \approx \arg \max_{t_1^n \in \text{Tagset}^n} \prod_{i=1}^n P(w_i | t_i) P(t_i | t_{i-1})$$

- Now we can easily estimate these two probabilities from a **part-of-speech tagged corpus**

$$P(t_i | t_{i-1}) = \frac{c(t_{i-1}, t_i)}{c(t_{i-1})}$$

$$P(w_i | t_i) = \frac{c(t_i, w_i)}{c(t_i)}$$

## Estimating Conditional Probabilities for Tags

$$P(t_i | t_{i-1}) = \frac{c(t_{i-1}, t_i)}{c(t_{i-1})}$$

- Examples:

$$P(NN | DT) = \frac{c(DT, NN)}{c(DT)} = \frac{58,800}{120,000} = 0.49$$

$$P(JJ | DT) = \frac{c(DT, JJ)}{c(DT)} = \frac{52,800}{120,000} = 0.42$$

$$P(IN | DT) = \frac{c(DT, IN)}{c(DT)} = \frac{120}{120,000} = 0.001$$

## Estimating Conditional Probabilities for Words

$$P(w_i | t_i) = \frac{c(t_i, w_i)}{c(t_i)}$$

- Examples:

$$P(is | VBZ) = \frac{c(VBZ, is)}{c(VBZ)} = \frac{9,600}{20,000} = 0.48$$

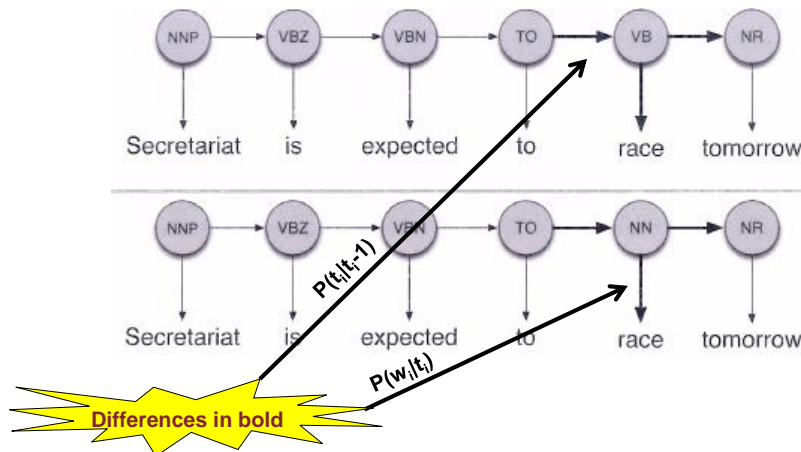
$$P(are | VBZ) = \frac{c(VBZ, are)}{c(VBZ)} = \frac{2}{20,000} = 0.0001$$

## An Example

- You can **book** your flight
  - P(book|VB)=0.0004
  - P(book|NN)=0.0002
  - P(VB|MD)=0.5
  - P(NN|MD)=0.001
- So what is the most likely tag for **book**?

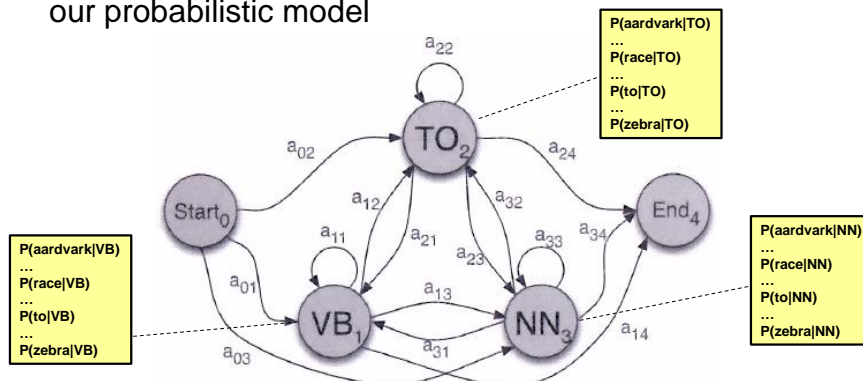
## Another Example from Jurafsky & Martin 2008

- How to choose the correct global tagging sequence?



## Hidden Markov Models (HMM)

- A HMM allows us to talk both about **observed events** (word sequence in input) and **unobserved (hidden) events** (part-of-speech tags) that are **causal factors** in our probabilistic model



Go to the Machine Learning class!!!



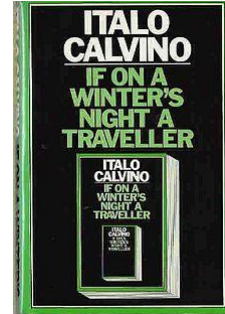
So far about word ordering...

- Morphological analysis: Finite-state transducers
- N-gram models: Computing probabilities for word sequences
- Part-of-speech classes: equivalence classes for words
- We now move to... formal grammars!



## Example

- Se una notte d'inverno un viaggiatore
- \*Se notte una d'inverno un viaggiatore
- Una notte se d'inverno un viaggiatore
- \*Se un notte d'inverno una viaggiatore
- Se una notte un viaggiatore d'inverno
- Se un viaggiatore d'inverno una notte
- \*Se un una notte viaggiatore d'inverno
- \*Se un una d'notte viaggiatore inverno
- ~Se un inverno d'notte un viaggiatore
- Se d'inverno un viaggiatore una notte
- Se d'inverno una notte un viaggiatore
- ...



## Context-Free Grammars (CFGs)

- A **context-free grammar (CFG)** or **phrase-structure grammar** is a formal grammar defined as a 4-tuple:

$$G = (N, T, P, S)$$

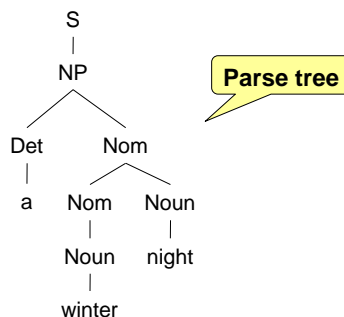
- where:
  - **N** is the set of nonterminal symbols (**phrases** or **clauses**)
  - **T** is the set of terminal symbols (**lexicon**)
  - **P** is the set of **productions (rules)**, a relation  $\subseteq N \times (N \cup T)^*$
  - **S** is the start symbol such that  $S \in N, \exists(S, \beta) \in P$

## Example

- $N = \{ S, NP, Nom, Det, Noun \}$ ,
- $T = \{ a, the, winter, night \}$ ,
- $P = \{$ 
  - $S \rightarrow NP$
  - $NP \rightarrow Det\ Nom$
  - $Nom \rightarrow Noun \mid Nom\ Noun$
  - $Det \rightarrow a \mid the$
  - $Noun \rightarrow winter$
  - $Noun \rightarrow night$ $\}$ ,
- $S$

## CFG as tools for...

- **Generating sentences**
  - G generates “a winter night”
  - There exists a **derivation** (sequence of rule expansions)



$N = \{ S, NP, Nom, Det, Noun \}$ ,  
 $T = \{ a, the, winter, night \}$ ,  
 $P = \{$ 

- $S \rightarrow NP$
- $NP \rightarrow Det\ Nom$
- $Nom \rightarrow Noun \mid Nom\ Noun$
- $Det \rightarrow a \mid the$
- $Noun \rightarrow winter$
- $Noun \rightarrow night$

 $\}$ .

- **Assigning a structure to a sentence**
  - What is the structure for “a winter night”?

## Treebanks

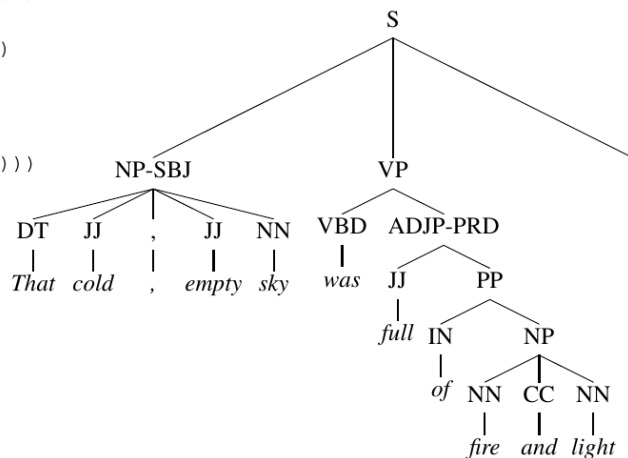


- CFGs can be used to **assign a parse tree** to any valid **sentence**
- We can build a corpus, called **treebank**, whose sentences are **annotated with parse trees**
- The most popular project of this kind is the **Penn Treebank**
  - From the **Brown**, **Switchboard**, **ATIS** and **Wall Street Journal** corpora of English
    - Wall Street Journal: 1.3 million words
    - Brown Corpus: 1 million words
    - Switchboard: 1 million words
  - All tagged with Part of Speech & syntactic structure
  - Developed 1988-1994

## “That cold, empty sky was full of fire and light.”



```
((S
  (NP-SBJ (DT That)
    (JJ cold) (, ,)
    (JJ empty) (NN sky) )
  (VP (VBD was)
    (ADJP-PRD (JJ full)
      (PP (IN of)
        (NP (NN fire)
          (CC and)
          (NN light) ))))
  (. .) ))
```



## Viewing Treebanks as Grammars



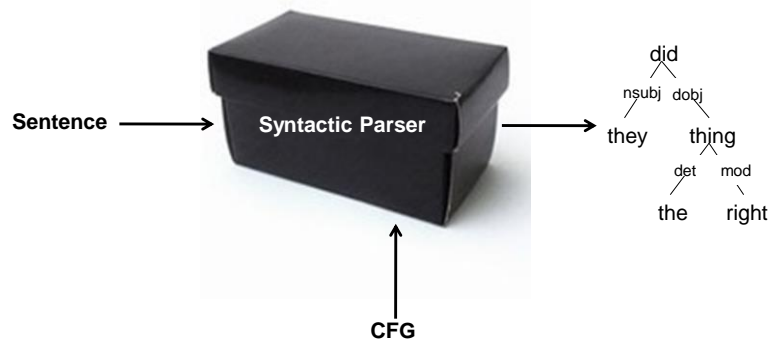
- The sentences in a treebank can be viewed as a **grammar of the language**
- We can extract the **rules** from the parsed sentences
- For example:
  - NP → DT JJ NN
  - NP → DT JJ NNS
  - NP → DT JJ NN NN **Cardinal number**
  - NP → DT JJ CD NNS
  - NP → RB DT JJ NN NN
  - NP → RB DT JJ JJ NNS
  - NP → DT JJ JJ NNP NNS

Adverb

Proper noun, sing.

## Syntactic Parsing

- The task of **recognizing a sentence** and assigning a **syntactic structure** to it



- However: CFGs do not specify how to calculate the **parse tree** for a sentence



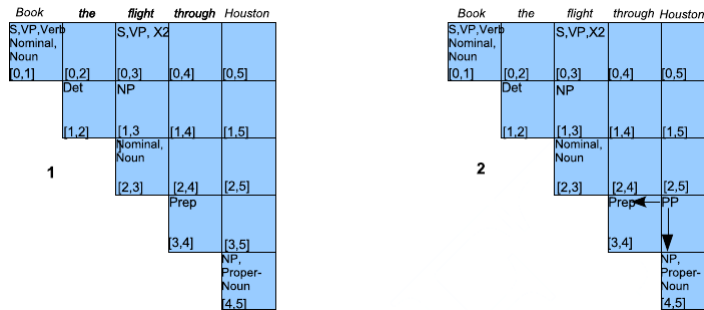
## The Cocke-Kasami-Younger (CKY) Algorithm

- A bottom-up dynamic programming parsing approach
- Takes as input a CFG in Chomsky Normal Form
- Given a sentence of  $n$  words, we need an  $(n+1) \times (n+1)$  matrix
- Cell  $(i,j)$  contains the set of non-terminals that produce all the constituents spanning positions from  $i$  to  $j$  of the input sentence
- The cell that represents the entire sentence is  $(0,n)$
- **Main idea:** if a non-terminal  $A$  is in  $(i,j)$ , there is a production  $A \rightarrow B C$ , so there must be an intermediate position  $k$  with  $B$  in  $(i,k)$  and  $C$  in  $(k,j)$

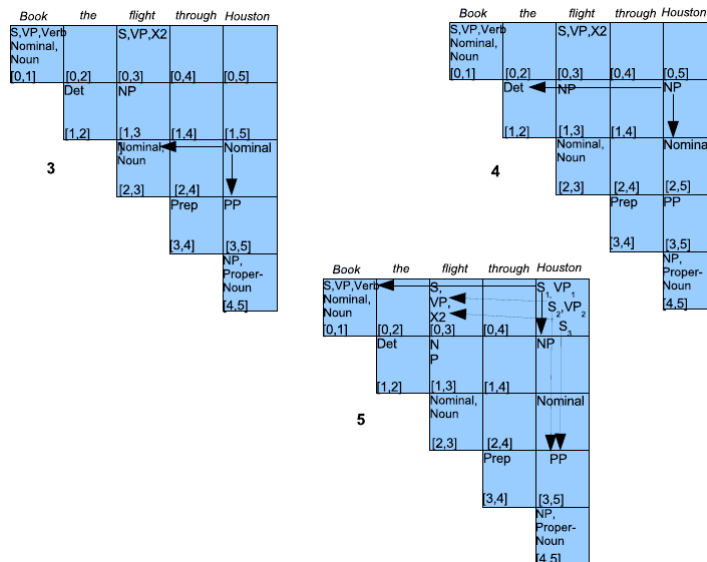
## Example of CKY Table [from Jurafsky & Martin book]

	<i>Book</i>	<i>the</i>	<i>flight</i>	<i>through</i>	<i>Houston</i>
S,VP,Verb Nominal, Noun			S,VP,X2		S, VP
[0,1]	[0,2]	[0,3]	[0,4]	[0,5]	
	Det	NP		NP	
	[1,2]	[1,3]	[1,4]	[1,5]	
		Nominal, Noun		Nominal	
		[2,3]	[2,4]	[2,5]	
			Prep	PP	
			[3,4]	[3,5]	
				NP, Proper- Noun	
				[4,5]	

## Example of CKY Table [from Jurafsky & Martin book]



## Example of CKY Table [from Jurafsky & Martin book]



## Probabilistic (or Stochastic) CFGs

- First proposed by Taylor Booth (1969)
- In a **probabilistic CFG**  $G = (N, T, P, S)$ , each production

$$A \rightarrow w [p]$$

is assigned a probability  $p = P(w|A) = P(A \rightarrow w)$

- For each left-hand-side non-terminal  $A$ , it must hold:

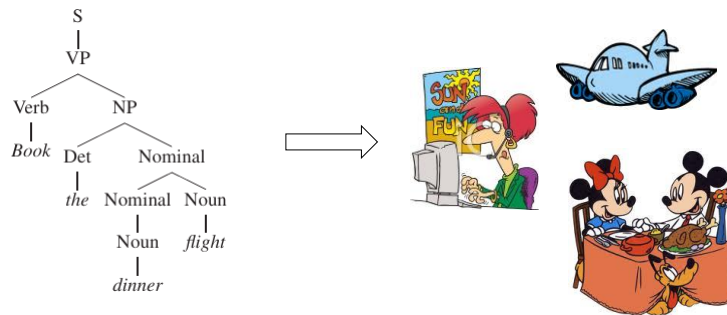
$$\sum_w P(A \rightarrow w) = 1$$

## An Example of PCFG (from Jurafsky & Martin)

$S \rightarrow NP VP$	[.80]	$Det \rightarrow that$	[.10]   $a$	[.30]   $the$	[.60]
$S \rightarrow Aux NP VP$	[.15]	$Noun \rightarrow book$	[.10]   $flight$	[.30]	
$S \rightarrow VP$	[.05]		$meal$	[.15]   $money$	[.05]
$NP \rightarrow Pronoun$	[.35]		$flights$	[.40]   $dinner$	[.10]
$NP \rightarrow Proper-Noun$	[.30]	$Verb \rightarrow book$	[.30]   $include$	[.30]	
$NP \rightarrow Det Nominal$	[.20]		$prefer$	[.40]	
$NP \rightarrow Nominal$	[.15]	$Pronoun \rightarrow I$	[.40]   $she$	[.05]	
$Nominal \rightarrow Noun$	[.75]		$me$	[.15]   $you$	[.40]
$Nominal \rightarrow Nominal Noun$	[.20]	$Proper-Noun \rightarrow Houston$	[.60]		
$Nominal \rightarrow Nominal PP$	[.05]		$TWA$	[.40]	
$VP \rightarrow Verb$	[.35]	$Aux \rightarrow does$	[.60]   $can$	[.40]	
$VP \rightarrow Verb NP$	[.20]	$Preposition \rightarrow from$	[.30]   $to$	[.30]	
$VP \rightarrow Verb NP PP$	[.10]		$on$	[.20]   $near$	[.15]
$VP \rightarrow Verb PP$	[.15]		$through$	[.05]	
$VP \rightarrow Verb NP NP$	[.05]				
$VP \rightarrow VP PP$	[.15]				
$PP \rightarrow Preposition NP$	[1.0]				

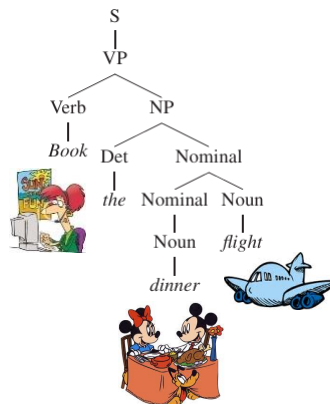
## Meaning, meaning, meaning!

- We are now moving from **syntax** to **semantics**



## Meaning, meaning, meaning!

- We are now moving from **syntax** to **semantics**



## Word Senses

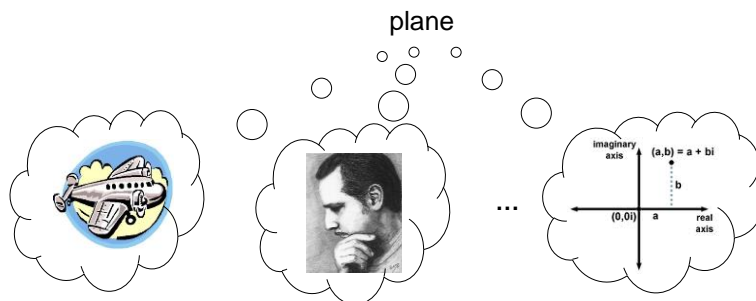
- The meaning of a word depends on the context in which it occurs



Context Matters

## Word Senses

- The meaning of a word depends on the context in which it occurs
- We call each meaning of a word a **word sense**



## Word Senses in Context

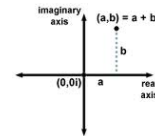
- I am catching the earliest **plane** to Brussels.



- This area probably lies more on the spiritual **plane** than the mental one.



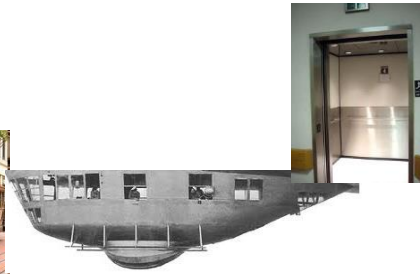
- Let's represent three-dimensional structures on a two-dimensional **plane**





## WordNet [Miller et al. 1990]

- The most popular computational lexicon of English
  - Based on psycholinguistic theories
- Concepts expressed as sets of synonyms (synsets)
  - {  $car_n^1$ ,  $auto_n^1$ ,  $automobile_n^1$ ,  $machine_n^4$ ,  $motorcar_n^1$  }
- A word sense is a word occurring in a synset
  - $machine_n^4$  is the fourth sense of noun machine

## WordNet: the “car” example

$$\begin{aligned} \text{Senses}_{WN}(\text{car}_n) = & \{ \{ \text{car}_n^1, \text{auto}_n^1, \text{automobile}_n^1, \text{machine}_n^4, \text{motorcar}_n^1 \}, \\ & \{ \text{car}_n^2, \text{rail car}_n^1, \text{rail way car}_n^1, \text{rail road car}_n^1 \}, \\ & \{ \text{cable car}_n^1, \text{car}_n^3 \}, \\ & \{ \text{car}_n^4, \text{gondola}_n^3 \}, \\ & \{ \text{car}_n^5, \text{elevator car}_n^1 \} \}. \end{aligned}$$


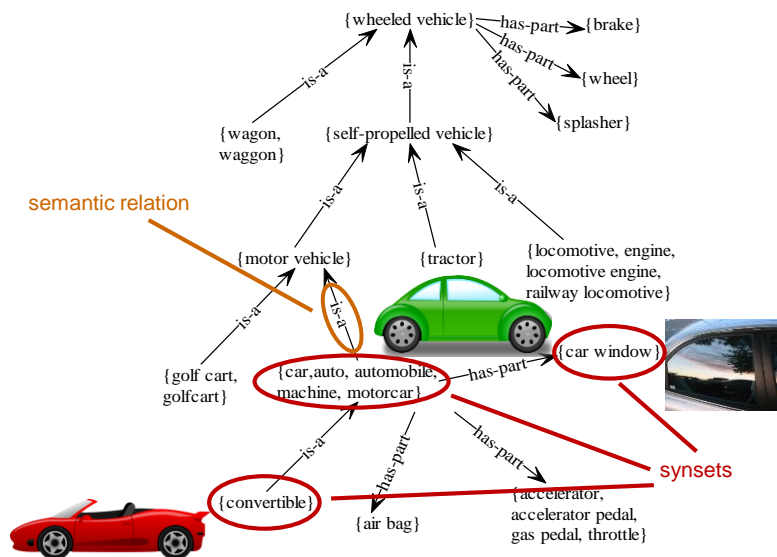
## WordNet provides textual definitions

- Called **glosses**
- A textual definition is provided for each synset
- Gloss of  $\text{car}_n^1$ : 
  - “a 4-wheeled motor vehicle; usually propelled by an internal combustion engine; ‘he needs a car to get to work’ ”
- Gloss of  $\text{car}_n^2$ : 
  - “a wheeled vehicle adapted to the rails of railroad; ‘three cars had jumped the rails’ ”
- Also available in **quasi-logical form**

## WordNet encodes relations!

- **Semantic relations** between synsets
  - **Hypernymy** (car<sub>n</sub><sup>1</sup> is-a motor vehicle<sub>n</sub><sup>1</sup>)
  - **Meronymy** (car<sub>n</sub><sup>1</sup> has-a car door<sub>n</sub><sup>1</sup>)
  - **Entailment, similarity, attribute**, etc.
- **Lexical relations** between word senses
  - **Synonymy** (i.e., words that belong to the same synset)
  - **Antonymy** (good<sub>a</sub><sup>1</sup> antonym of bad<sub>a</sub><sup>1</sup>)
  - **Pertainymy** (dental<sub>a</sub><sup>1</sup> pertains to tooth<sub>n</sub><sup>1</sup>)
  - **Nominalization** (service<sub>n</sub><sup>2</sup> nominalizes serve<sub>v</sub><sup>4</sup>)

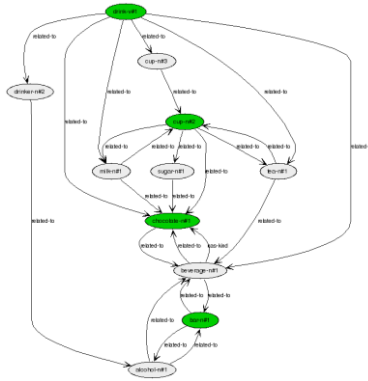
## WordNet as a Graph





## But WordNet is more than Simply a Graph!

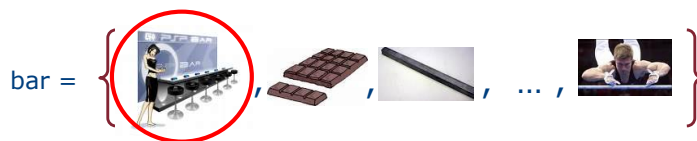
- It is a semantic network!
- A semantic network is a network which represents semantic relations among concepts
- It is often used as a form of knowledge representation



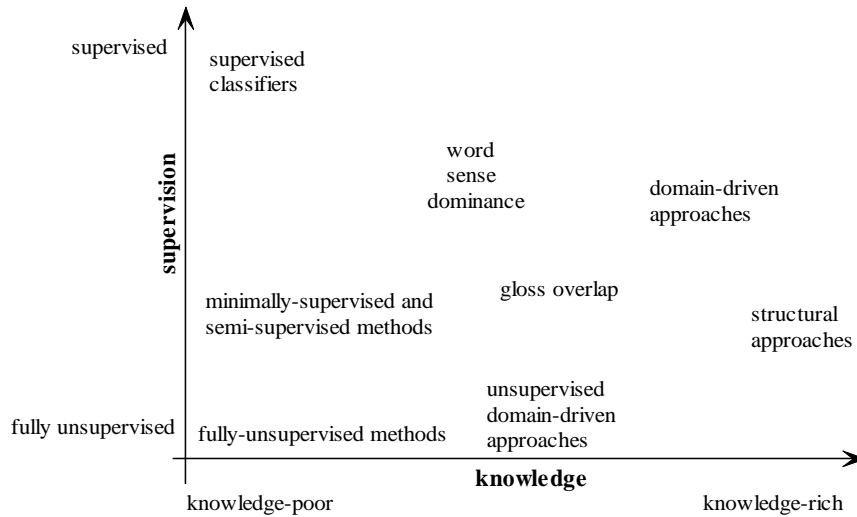
## Word Sense Disambiguation (WSD)

- WSD is the task of computationally determining which sense of a word is activated by its use in a particular context [Ide and Véronis, 1998; Navigli, 2009]
- It is basically a classification task
  - The objective is to learn how to classify words into word senses
  - This task is strongly tied to Machine Learning

I drank a cup of chocolate at the **bar**

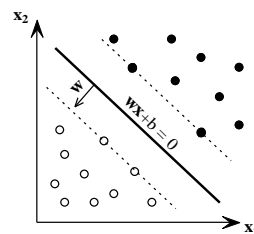


## Supervision and Knowledge

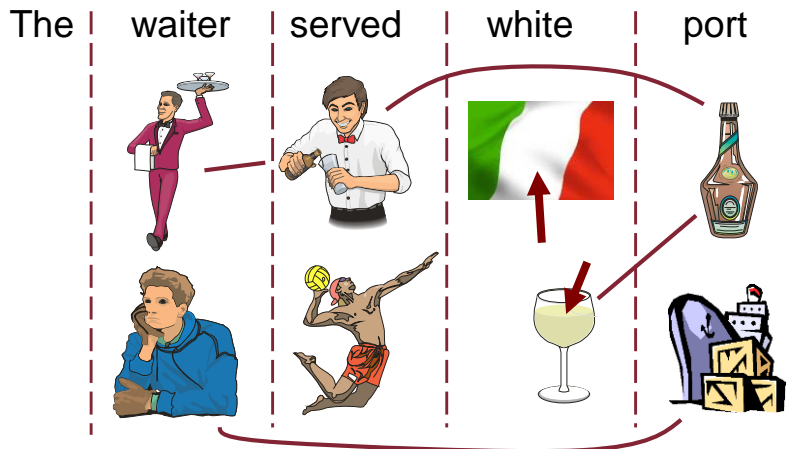


## Supervised WSD: Support Vector Machines

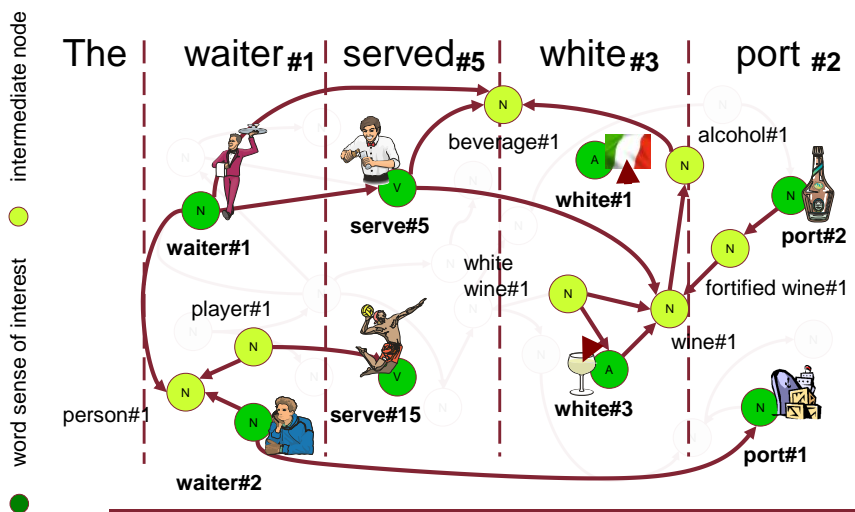
- SVM learns a **linear hyperplane** from the training set that separates positive from negative examples
- The hyperplane maximizes the distance to the closest positive and negative examples (**support vectors**)
- Achieves state-of-the-art performance in WSD [Keok and Ng, 2002]



## Knowledge-based Word Sense Disambiguation

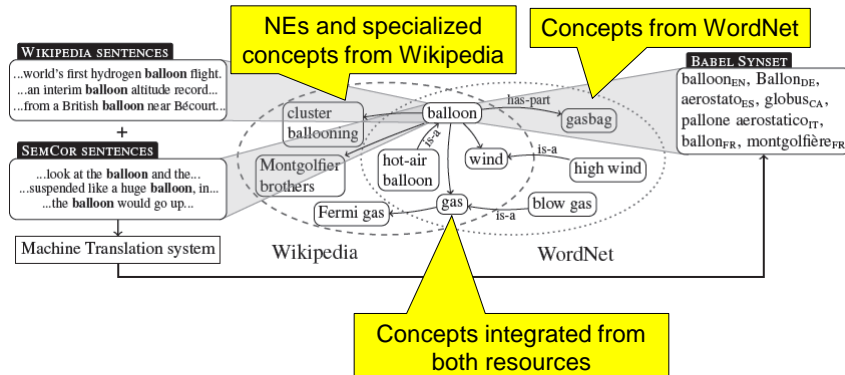


## Knowledge-based Word Sense Disambiguation



## BabelNet [Navigli and Ponzetto, AIJ 2012]

- A wide-coverage multilingual semantic network including both **encyclopedic** (from Wikipedia) and **lexicographic** (from WordNet) entries



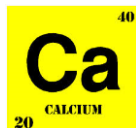
BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

129

## BabelNet as a Multilingual Inventory for:

### Concepts

*Calcio* in Italian can denote different concepts:



### Named Entities

The word *Mario* can be used to represent different things such as the video game character or a soccer player (Gomez) or even a music album



BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

130

BabelNet 3.0 is online: <http://babelnet.org>



A very large multilingual encyclopedic dictionary and semantic network

Type a text or a term...

ENGLISH

SEARCH

⚙️ PREFERENCES

BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

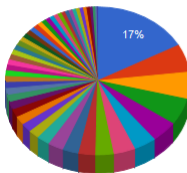
131

## Anatomy of BabelNet 3.0

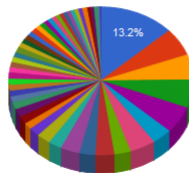
- **271 languages covered** (including Latin!)
- **13.8M Babel synsets**
  - (6.4M concepts and 7.4M named entities)
- **117M word senses**
- **355M semantic relations** (26 edges per synset on avg.)
- **11M synset-associated images**
- **40M textual definitions**



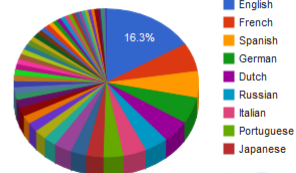
Lemmas by Language



Synsets by Language



Senses by Language



BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

132

## New 3.0 version out!

- Seamless integration of:
  - **WordNet 3.0**
  - **Wikipedia**
  - **Wikidata**
  - **Wiktionary**
  - **OmegaWiki**
  - **Open Multilingual WordNet** [Bond and Foster, 2013]
- Translations for **all open-class parts of speech**
- **2B** RDF triples available via SPARQL endpoint



BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

133

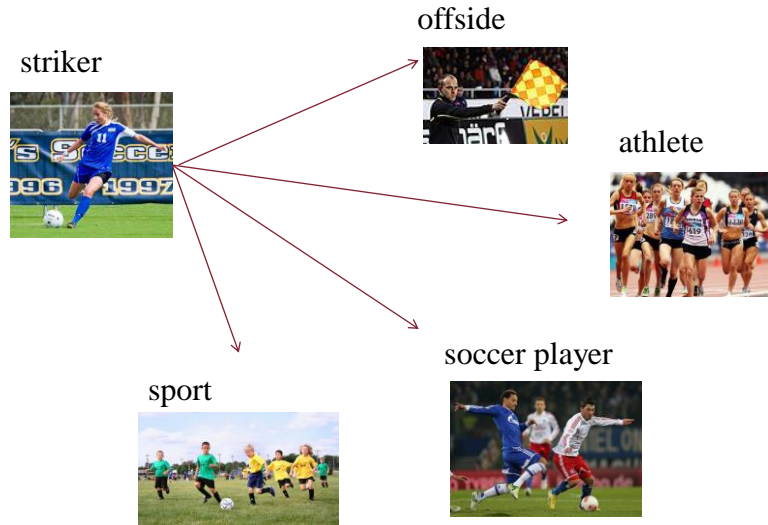
## So what?



BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

134

## Step 1: Semantic Signatures



BabelNet, Babelify, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

135

## Step 2: Find all possible meanings of words

### 1. Exact Matching (good for WSD, bad for EL)

~~Thomas~~ and Mario are ~~s~~  ~~rs~~ playing in Munich



Thomas,  
Norman



Thomas,  
Seth



They both have  
Thomas as one of  
their lexicalizations

BabelNet, Babelify, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

137

## Step 2: Find all possible meanings of words

### 2. Partial Matching (good for EL)

Thomas and Mario are strikers playing in Munich



Thomas,  
Norman



Thomas,  
Seth



Thomas  
Müller

It has Thomas as a substring of one of its lexicalizations

BabelNet, Babelify, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

138

## Step 2: Find all possible meanings of words

- “Thomas and Mario are strikers playing in Munich”

Seth Thomas



Mario (Character)



striker (Sport)



Munich (City)



Thomas Müller



Mario (Album)



Striker (Video Game)



FC Bayern Munich



Thomas (novel)



Mario Gómez



Striker (Movie)



Munich (Song)



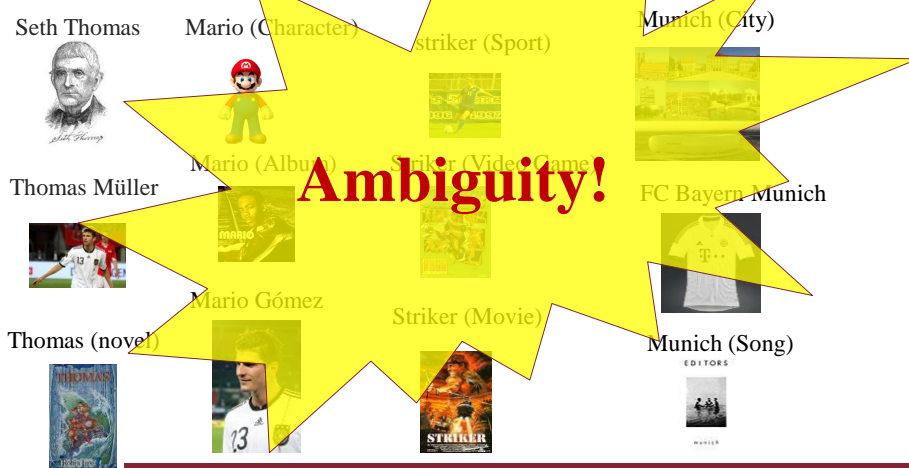
BabelNet, Babelify, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

139



## Step 2: Find all possible meanings of words

- “Thomas and Mario are strikers playing in Munich”

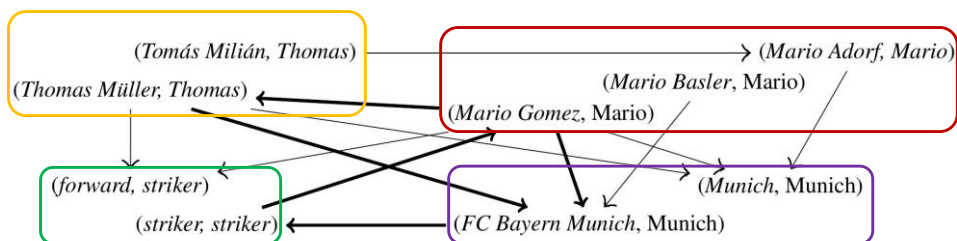


BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

140

## Step 3: Connect all the candidate meanings

- **Thomas** and **Mario** are **strikers** playing in **Munich**

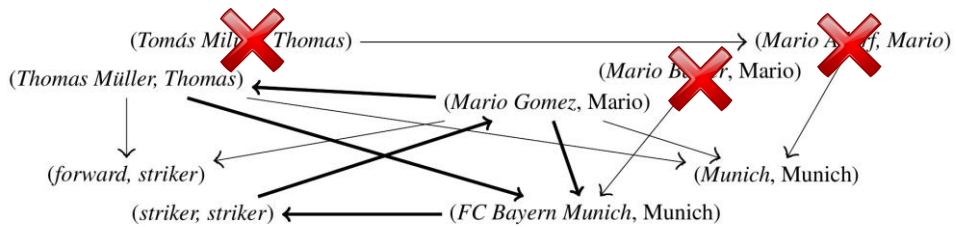


BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

142

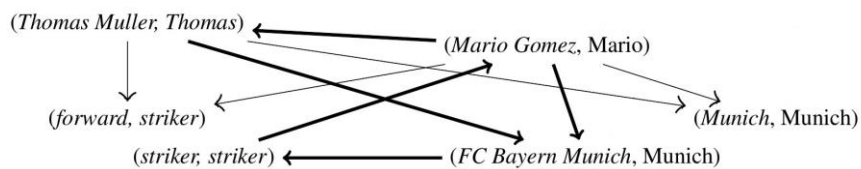
#### Step 4: Extract a dense subgraph

- **Thomas** and **Mario** are **strikers** playing in **Munich**



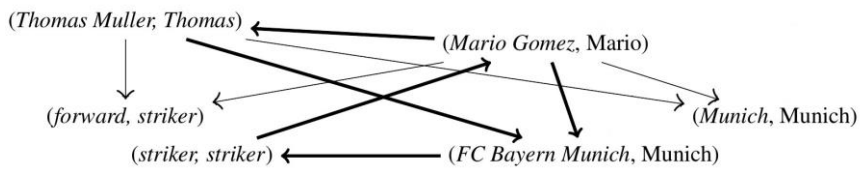
#### Step 4: Extract a dense subgraph

- **Thomas** and **Mario** are **strikers** playing in **Munich**



## Step 5: Select the most reliable meanings

- **Thomas and Mario are strikers playing in Munich**




## Step 5: Select the most reliable meanings

- “Thomas and Mario are strikers playing in Munich”



<http://babelfy.org>



fy

Babelfy

[Text to babelfy...]




Enable partial matches:

ENGLISH

BABELFY!



ABOUT  
PUBLICATIONS  
DOWNLOADS


Babelfy is an output of the [MultiJEDI ERC Starting Grant](#) No. 259234. Concept and application by [Roberto Navigli](#). Babelfy and its API are licensed under a [Creative Commons Attribution-Non Commercial-Share Alike 3.0 License](#). For any commercial use, please [contact us](#).   



BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

150

## The Charlie Hebdo gun attack (English)



fy

Babelfy

Charlie ~~Hebdo~~: Gun attack on French magazine kills 12. Gunmen have shot dead 12 people at the Paris office of French satirical magazine Charlie ~~Hebdo~~ in an apparent militant Islamist attack.

Four of the magazine's well-known cartoonists, including its editor, were among those killed, as well as two police officers.


Enable partial matches:

ENGLISH


BABELFY!

expanded view | compact view


Charlie Hebdo : Gun attack on French magazine kills 12 .




**Charlie Hebdo**  
Charlie Hebdo is a French satirical weekly newspaper, featuring cartoons.




**Gun**  
a weapon that discharges a missile at high velocity (especially from a



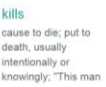
**attack**  
(military) an offensive against an enemy (using weapons); "the attack began at dawn"



**French**  
of or pertaining to France or the people of France; "French cooking"; "a Gallic



**magazine**  
a periodic publication containing pictures and stories and articles of interest to



**kills**  
cause to die; put to death, usually intentionally or knowingly; "This man

BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

## The Charlie Hebdo gun attack (English)

The screenshot shows the BabelFY search interface. At the top, the search results for 'Charlie Hebdo' are displayed, including a snippet: 'Gun attack on French magazine kills 12. Gunmen have shot dead 12 people at the Paris office of French satirical magazine Charlie Hebdo in an apparent militant Islamist attack.' Below this, there are controls for 'Enable partial matches' and a language dropdown set to 'ENGLISH'. The main part of the interface is a word cloud where the words 'Gunmen', 'shot', 'dead', '12', 'people', 'at the', 'Paris', and 'offi' are highlighted. Below the word cloud, there are several cards with images and definitions for the highlighted words: 'Gunmen' (a professional killer who uses a gun), 'shot' (hit with a missile from a weapon), 'dead' (quickly and without warning; "he stopped suddenly"), 'people' (plural) any group of human beings (men or women or children) collectively; "old", and 'Paris' (the capital and largest city of France; and international center of culture and commerce).

BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

## The Charlie Hebdo gun attack (English)

This screenshot shows the BabelFY search interface with a different word cloud. The search results for 'Charlie Hebdo' are the same as in the first screenshot. The word cloud highlights the words 'office', 'of', 'French', 'satirical', 'magazine', 'Charlie Hebdo', 'in an', and 'apparent'. Below the word cloud, there are several cards with images and definitions for the highlighted words: 'office' (an administrative unit of government; "the Central Intelligence Agency"; "the Census Bureau"), 'French' (of or pertaining to France or the people of France; "French cooking"; "a Gallic"), 'satirical' (exposing human folly to ridicule; "a persistent campaign of mockery by the press"), 'magazine' (a periodic publication containing pictures and stories and articles of interest to a particular group), 'Charlie Hebdo' (Charlie Hebdo is a French satirical weekly newspaper, featuring cartoons, and known for its provocative content), and 'apparent' (clearly revealed to the mind or the senses or judgment; "the effects of the drought are apparent").

BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

## The Charlie Hebdo gun attack (Italian)

The screenshot shows the Babelfy interface with a search bar containing the text "Assalto al giornale Charlie Hebdo: 12 morti, Due dei tre killer reduci dalla Siria". Below the search bar, there are several cards representing search results:

- Assalto**: close fighting during the culmination of a military attack
- Charlie Hebdo**: Charlie Hebdo is a French satirical weekly newspaper, featuring cartoons.
- morti**: the event of dying or departure from life; "Her death came as a terrible shock"; "upon
- Due**: being one more than one; "he received two messages"
- tre**: being one more than two

BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

## The Charlie Hebdo gun attack (Italian)

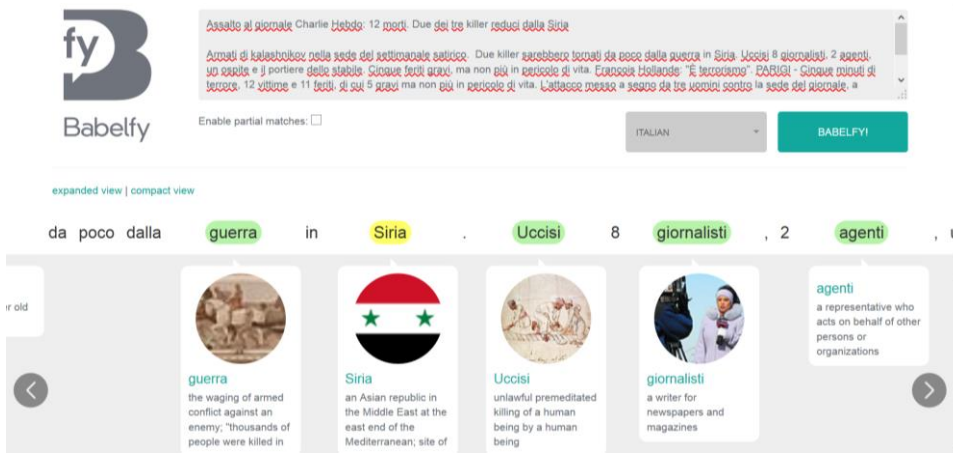
The screenshot shows the Babelfy interface with a search bar containing the text "Assalto al giornale Charlie Hebdo: 12 morti, Due dei tre killer reduci dalla Siria". Below the search bar, there are several cards representing search results:

- tre**: being one more than two
- killer**: a professional killer who uses a gun
- reduci dalla**: a type of submachine gun made in Russia
- Siria**: an Asian republic in the Middle East at the east end of the Mediterranean; site of
- Armati**: an enlisted man or woman who serves in an army; "the soldiers stood at attention"
- di**: a type of submachine gun made in Russia
- kalashnikov**: a type of submachine gun made in Russia
- nella**: (plural) a military unit consisting of a commander and the headquarters staff
- sede**: (plural) a military unit consisting of a commander and the headquarters staff

BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

155

## The Charlie Hebdo gun attack (Italian)



Assalto al giornale Charlie Hebdo: 12 morti. Due dei tre killer reduci dalla Siria.

Armati di kalashnikov nella sede del settimanale satirico. Due killer sarebbero tornati da poco dalla guerra in Siria. Uccisi 8 giornalisti, 2 agenti. Un quarto è il portiere dello stabile. Si vuole farci capire, ma non più in agguato di vita. European headline: "È l'attacco". BABELFY - Giocare parole di guerra: 12 vittime e 11 feriti, di cui 8 gravi ma non più in pericolo di vita. L'attacco messo a segno da tre uomini scappati la sede del giornale.

Enable partial matches:

ITALIAN BABELFY!

expanded view | compact view

da poco dalla guerra in Siria, Uccisi 8 giornalisti, 2 agenti

guerra  
the waging of armed conflict against an enemy; "thousands of people were killed in

Siria  
an Asian republic in the Middle East at the east end of the Mediterranean; site of

Uccisi  
unlawful premeditated killing of a human being by a human being

giornalisti  
a writer for newspapers and magazines

agenti  
a representative who acts on behalf of other persons or organizations

BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

## The Charlie Hebdo gun attack (French)



Charlie Hebdo visé par une attaque terroriste, deux national décapité. La rédaction du journal satirique Charlie Hebdo a été la cible d'un attentat meurtrier 7 heures en fin de journée à Paris, qui a fait douze morts. Un avis de recherche a été lancé en début de nuit contre trois suspects.

Le plus jeune des auteurs ressemblait à un 18 ans, sans doute dans le rôle du commissariat de Charbonnières-les-Bains dans les Alpes. Une source policière, consultée par Le Monde, a affirmé qu'« aucune charge » n'avait été retenue pour l'un des auteurs du attentat, et que « dans son cas, il ne

Enable partial matches:

FRENCH BABELFY!

expanded view | compact view

Paris, qui a fait douze morts. Un avis de recherche a été lancé en déb

fait  
behave in a certain manner; show a certain behavior; conduct or comport

morts  
the event of dying or departure from life; "her death came as a terrible shock"; "upon

avis de recherche  
a public announcement by a law enforcement

lancé  
carry out or perform an action; "John did the painting, the wedding, and he

début  
the event of the start of "the beginning war"

BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

## The Charlie Hebdo gun attack (French)

fy Babelfy

Charlie Hebdo visé par une attaque terroriste: deux national décapités. La rédaction du journal satirique Charlie Hebdo a été le cible d'un attentat meurtrier 7 heures en fin de nuit à Paris, qui a fait deux morts. Au cours de recherche a été lancé en début de nuit contre trois suspects.

Le plus jeune homme recherché, âgé de 18 ans, s'est rendu dans la nuit au commissariat de Charleville-Mézières dans les Ardennes. Une source policière, citée par Le Monde a affirmé qu'« aucune charge » n'avait été retenue pour l'homme contre lui, et que « dans son cas, il de

Enable partial matches:

FRENCH BABELFY!

expanded view | compact view

en début de nuit contre trois suspects . Le plus jeune homme re

rm did

début the event consisting of the start of something; "the beginning of the war"

nuit the time after sunset and before sunrise while it is dark outside

suspects a person or institution against whom an action is brought in a court of law; the

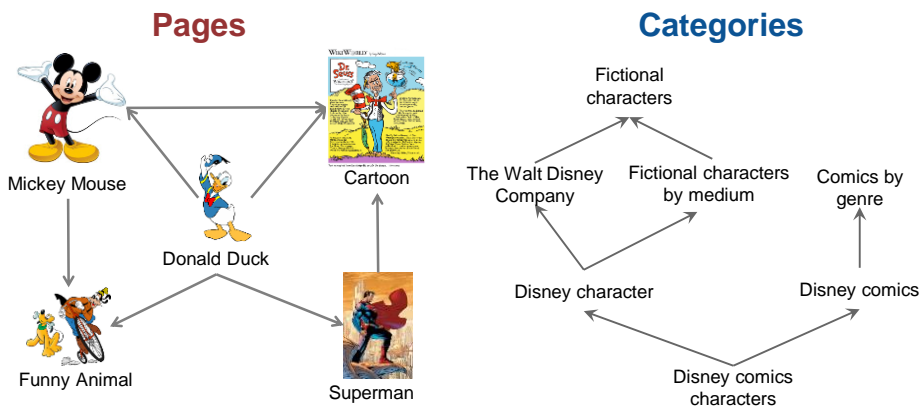
plus advancing in amount or intensity; "she became increasingly depressed"

jeune homme a youthful male person; "the baby was a boy"; "she made the boy brush his teeth"

rech look f persc

BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

## The Wikipedia structure: an example



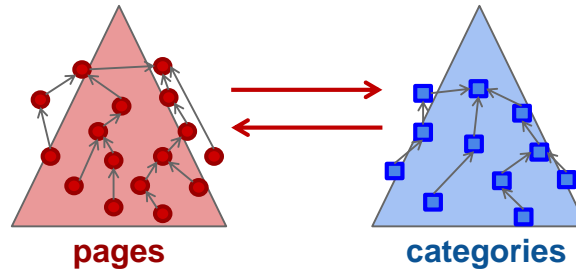
BabelNet, Babelfy, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

159



## Our goal

To **automatically** create a **Wikipedia Bitaxonomy** for Wikipedia **pages** and **categories** in a simultaneous fashion.



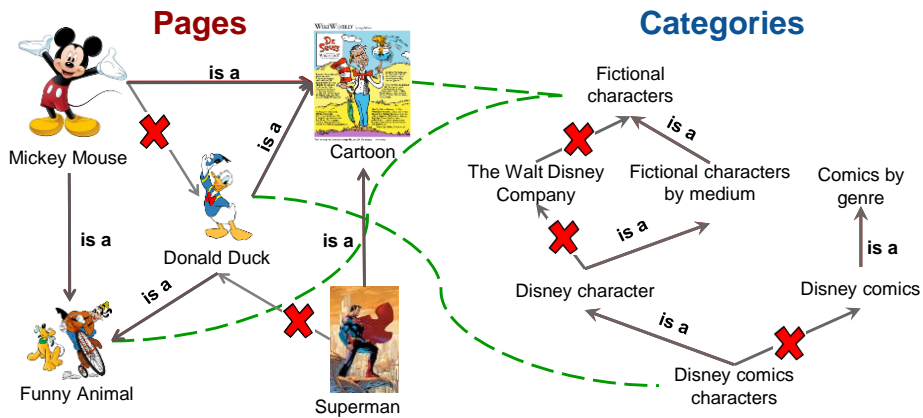
## Our goal

To **automatically** create a **Wikipedia Bitaxonomy** for Wikipedia **pages** and **categories** in a simultaneous fashion.

### KEY IDEA

The **page** and **category** level are **mutually beneficial** for inducing a **wide-coverage** and **fine-grained** integrated taxonomy

## The Wikipedia Bitaxonomy: an example



BabelNet, Babelify, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

162



## The WiBi Page taxonomy

## Assumption

- The **first sentence** of a page is a good **definition** (also called gloss)



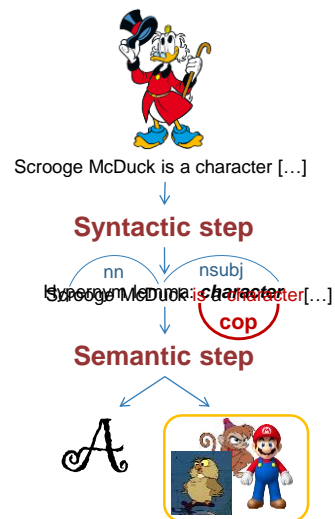
### Scrooge McDuck

From Wikipedia, the free encyclopedia

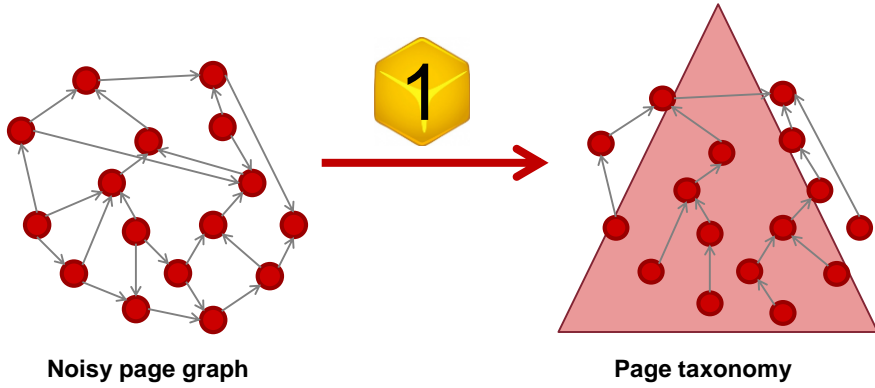
**Scrooge McDuck** is a cartoon character created in 1947 by Carl Barks and licensed by The Walt Disney Company. Scrooge is an elderly Scottish anthropomorphic white duck with a yellow-orange bill, legs, and feet.

## The WiBi Page taxonomy

1. **[Syntactic step]**  
Extract the hypernym lemma from a page definition using a syntactic parser;
2. **[Semantic step]**  
Apply a set of linking heuristics to disambiguate the extracted lemma.



## The story so far



## The Bitaxonomy algorithm

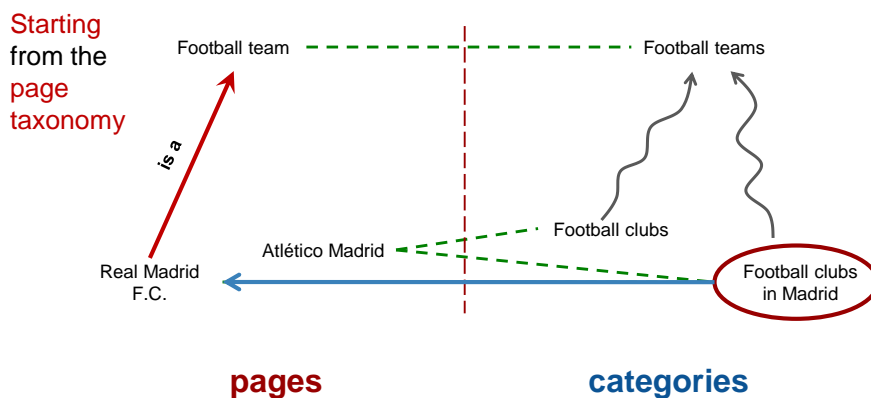
## The Bitaxonomy algorithm

The information available in the two taxonomies is **mutually beneficial**

- At each step **exploit one taxonomy to update the other** and vice versa
- Repeat until **convergence**

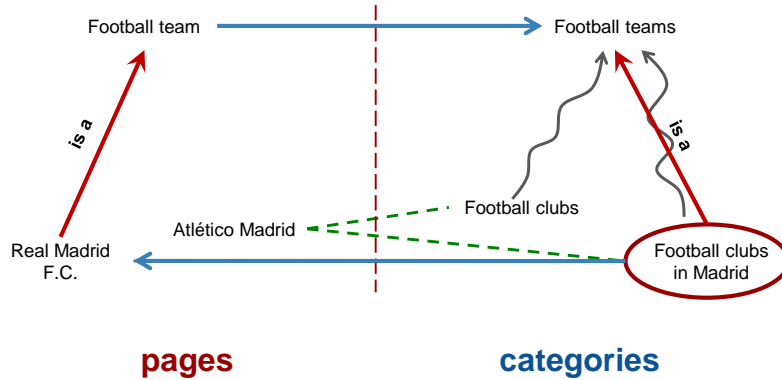


## The Bitaxonomy algorithm



## The Bitaxonomy algorithm

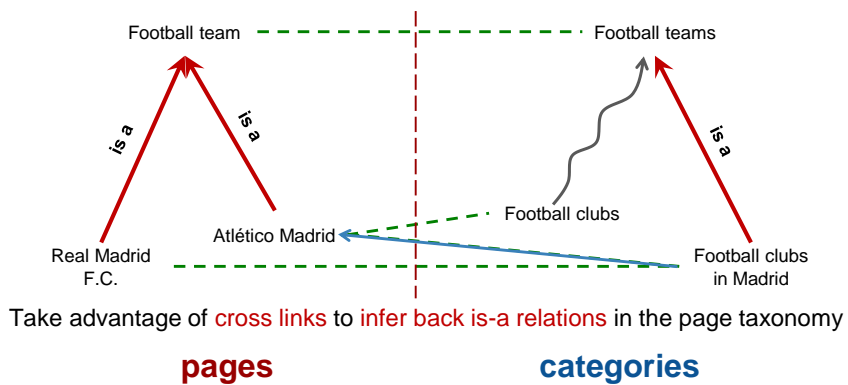
Exploit the **cross links** to infer **hypernym** relations in the category taxonomy



BabelNet, Babelify, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

170

## The Bitaxonomy algorithm

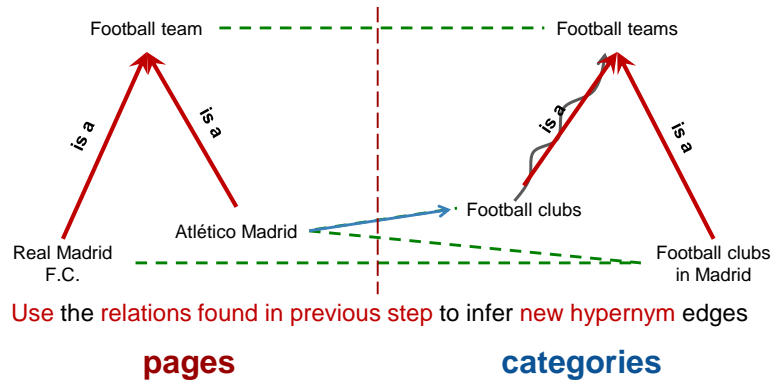


Take advantage of **cross links** to infer **back is-a relations** in the page taxonomy

BabelNet, Babelify, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

171

## The Bitaxonomy algorithm

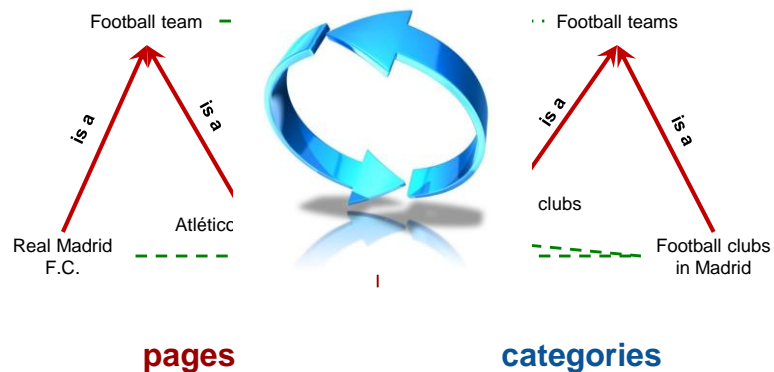


BabelNet, Babelify, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

172

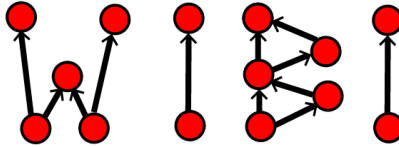
## The Bitaxonomy algorithm

Mutual enrichment of both taxonomies until convergence



BabelNet, Babelify, Video games with a purpose & the Wikipedia Bitaxonomy  
Roberto Navigli

173



WiBi (Wikipedia Bitaxonomy) is an approach to the automatic creation of a bitaxonomy for Wikipedia developed by Tiziano Flati, Daniele Vannella, Tommaso Pasini, and Roberto Navigli.

WiBi is now also integrated into BabelNet

Input a Wikipedia item

Search

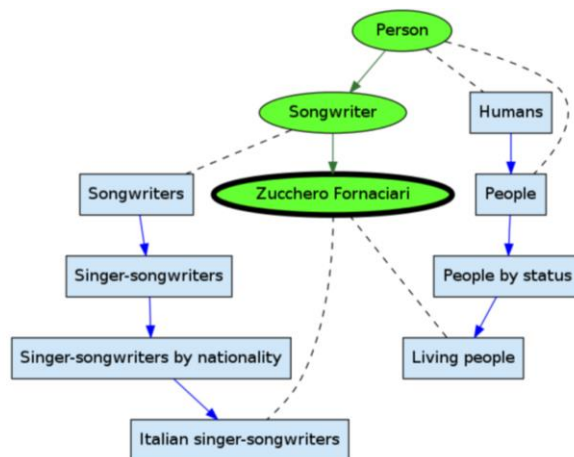
Try out some examples:

The Da Vinci Code (film), Zucchero Fornaciari, Różynek Wielki, Moulin Rouge, WordNet, Julia Roberts, Florence, ABBA, Eric Nies, Mąkosy Stare



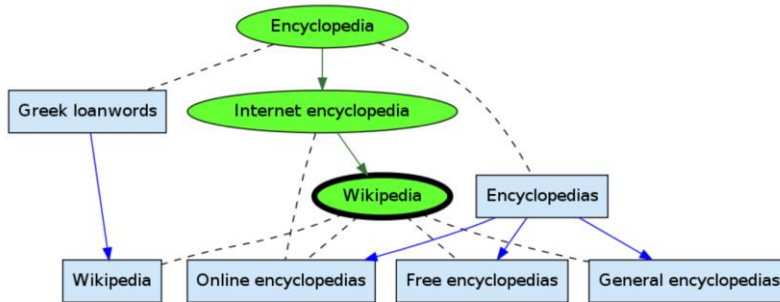
wibitaxonomy.org

### Example from <http://wibitaxonomy.org>: WordNet



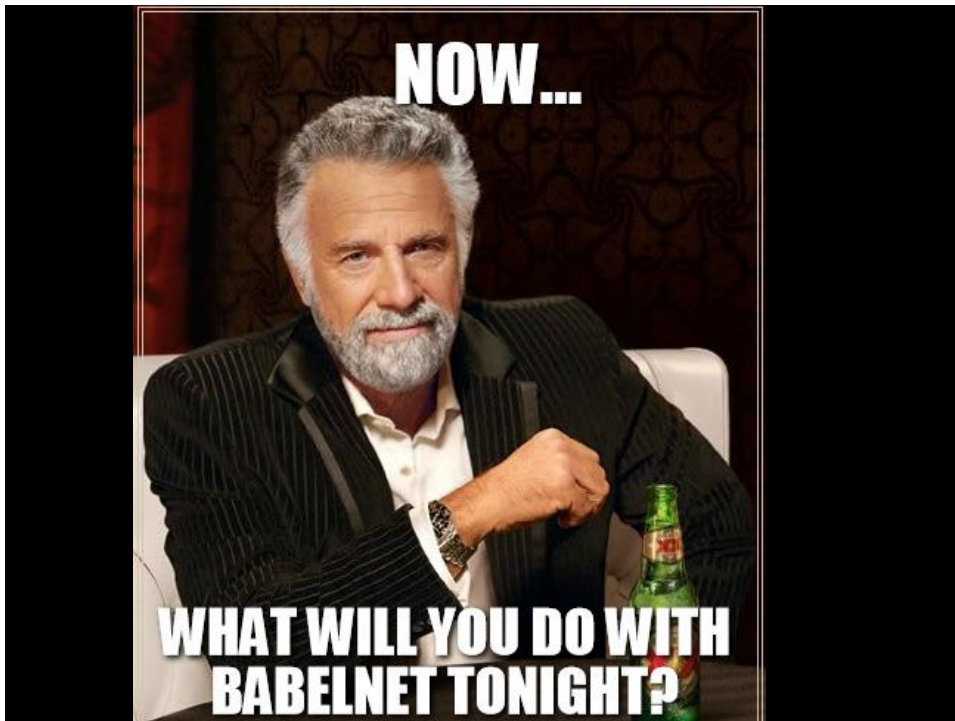


Example from <http://wibitaxonomy.org>: Wikipedia



**THAT'S ALL FOLKS!!!**

(Isn't it enough???)



Thanks or...





SAPIENZA  
UNIVERSITÀ DI ROMA

**Roberto Navigli**

Linguistic Computing Laboratory  
<http://lcl.uniroma1.it>

