

# Babelplagiarism: what can BabelNet do for cross-language plagiarism detection?

Roberto Navigli

DIPARTIMENTO  
DI INFORMATICA



SAPIENZA  
UNIVERSITÀ DI ROMA

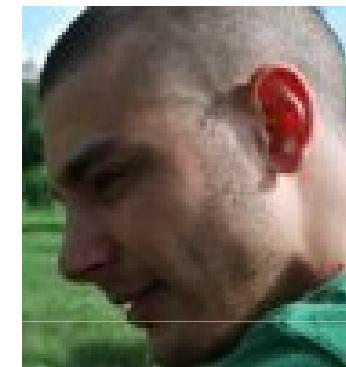
## Joint work with...



Simone Ponzetto



Mirella Lapata



Andrea Moro

# Outline

- Motivation: the knowledge acquisition bottleneck
- BabelNet: constructing a large-scale multilingual ontology
- What can BabelNet do for (cross-language) plagiarism detection?
- Conclusions: lessons learned

# It's all about knowledge!

- Intuitively, we all **know** what **knowledge** is...
- ...and why we need it
- But can we expect computers to **know**?
- Can't computers just use, e.g., **statistical techniques**?

analyze and compare the mix of activities performed in terms of the available knowledge bases proprietary knowledge used peers as a measure of understanding commonalities

# Machine Translation (Google Translate)

EN: I love chocolate, so I bought a bar in the supermarket.



# Machine Translation (Google Translate)

- EN: These are movies in which the music genre, e.g. **rock**, is an important element but not necessarily central to the plot. Examples are Easy Rider (1969), The Graduate (1969), and Saturday Night Fever (1978).



# Machine Translation (Google Translate)

- **EN:** These are movies in which the music genre, e.g. **rock**, is an important element but not necessarily central to the plot. Examples are Easy Rider (1969), The Graduate (1969), and Saturday Night Fever (1978).
- **IT:** Questi sono i film in cui il genere musicale, ad es **roccia**, è un elemento importante, ma non necessariamente al centro della trama.



# Machine Translation (Google Translate)

- EN: Knowledge of the distribution of underground **rock** densities can assist in interpreting subsurface geologic structure and rock type.



# Machine Translation (Google Translate)

- EN: Knowledge of the distribution of underground **rock** densities can assist in interpreting subsurface geologic structure and rock type.
- IT: La conoscenza della distribuzione di densità di **rock underground** può aiutare a interpretare in sottosuolo struttura geologica e tipo di roccia.



It's not that the “big data” approach is bad,  
it's just that mere statistics is not enough



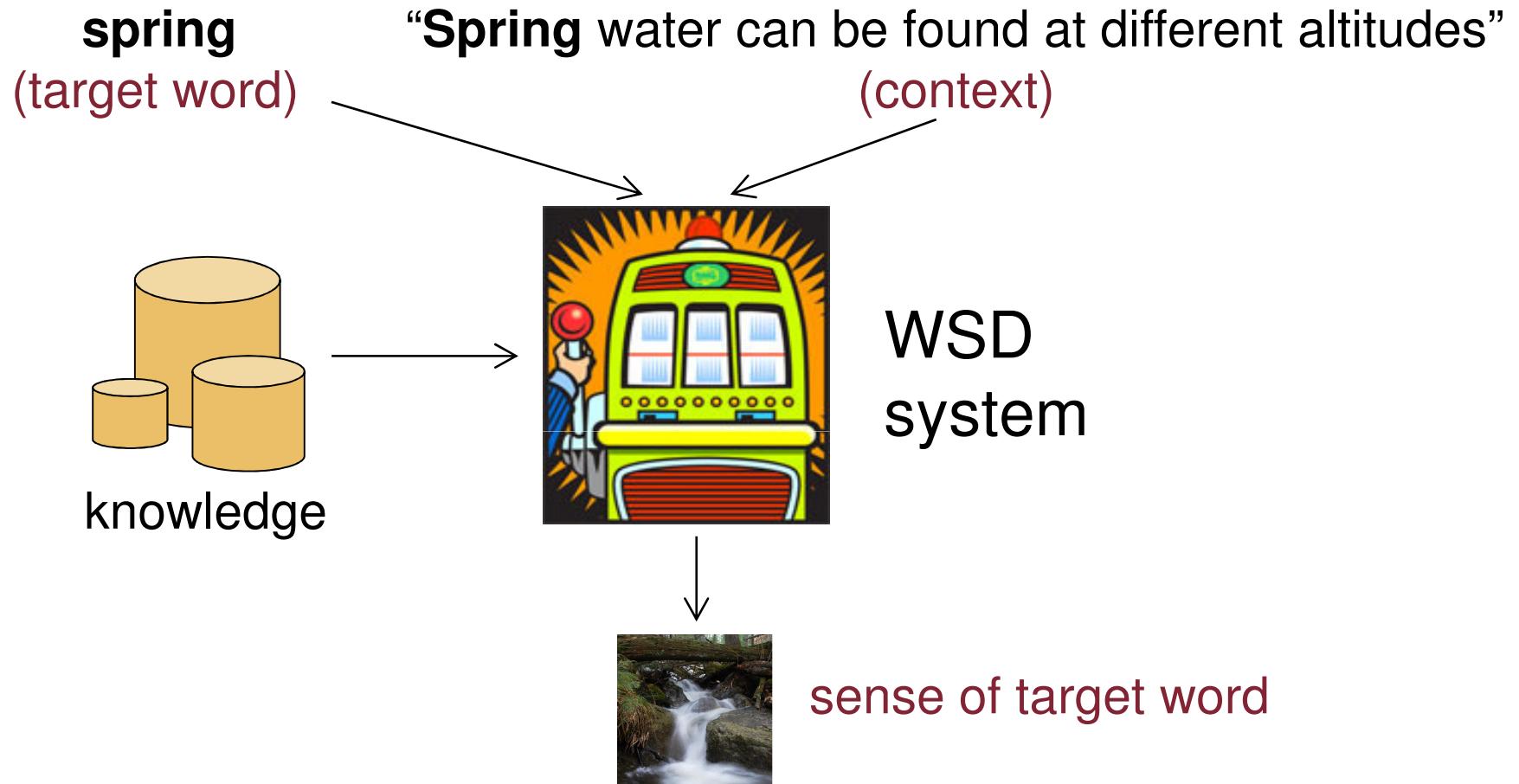
# The Knowledge Acquisition Bottleneck

- Knowledge is crucial in NLP
  - Word Sense Disambiguation
  - Named Entity Recognition
  - Question Answering
  - (your favourite NLP task here)
- However, providing knowledge is difficult and costly
- Various projects undertaken to make lexical knowledge available in a machine readable format
  - WordNet [Fellbaum, 1998]
  - Open Mind Word Expert [Chklovski & Mihalcea, 2002]
  - The WordNetPlus project [Boyd-Graber et al., 2006]
  - OntoNotes [Hovy et al., 2006]
  - EuroWordNet [Vossen, 1998], Multilingual Central Repository [Atserias et al. 2004], ...
  - Wikipedia (collaborative effort)



Plagiarism detection!

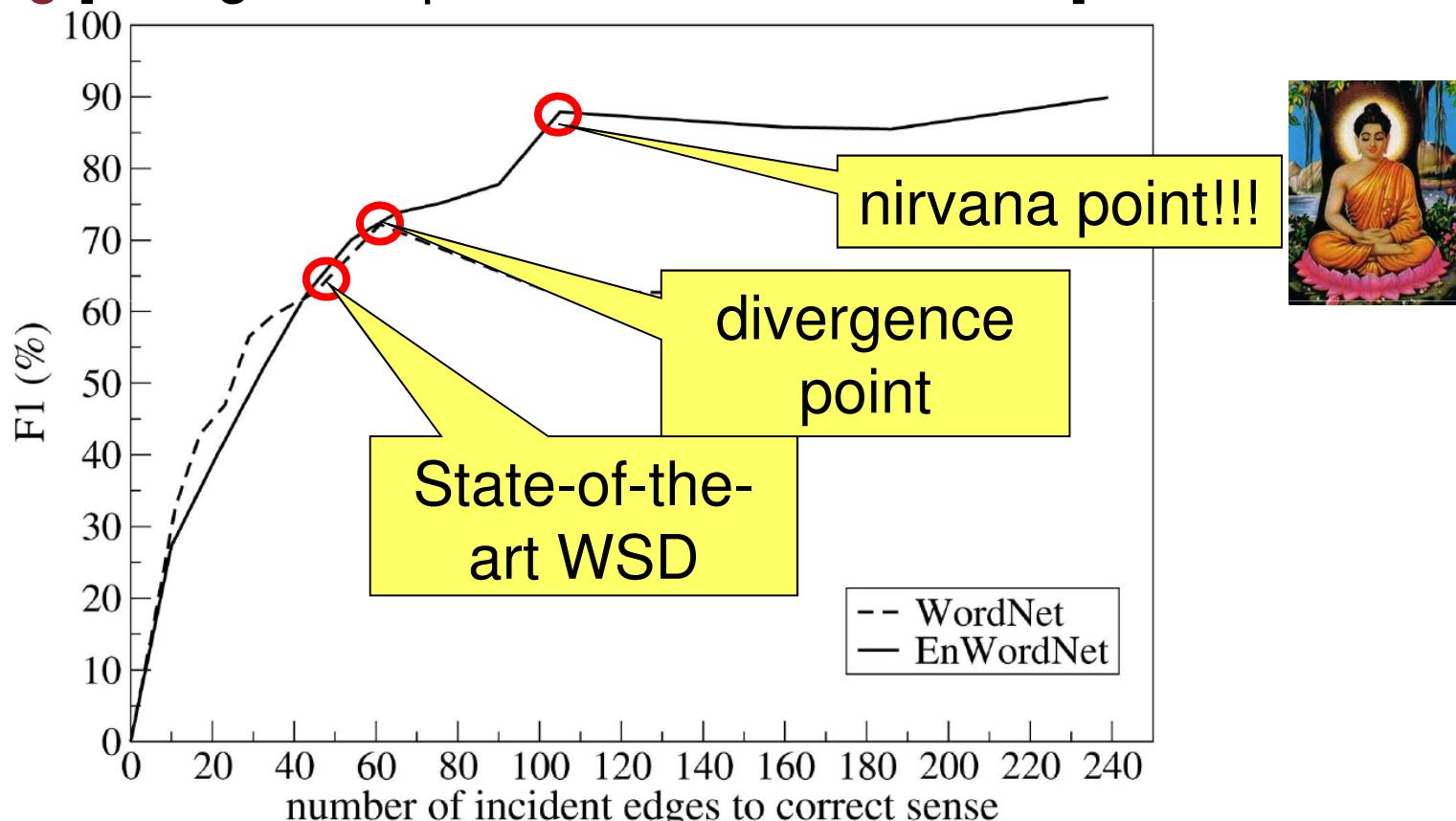
# Word Sense Disambiguation in a Nutshell



Roberto Navigli: Word sense disambiguation: A survey. ACM Computing Surveys 41(2), 2009, pp. 1-69

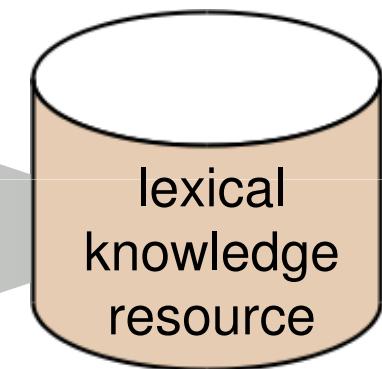
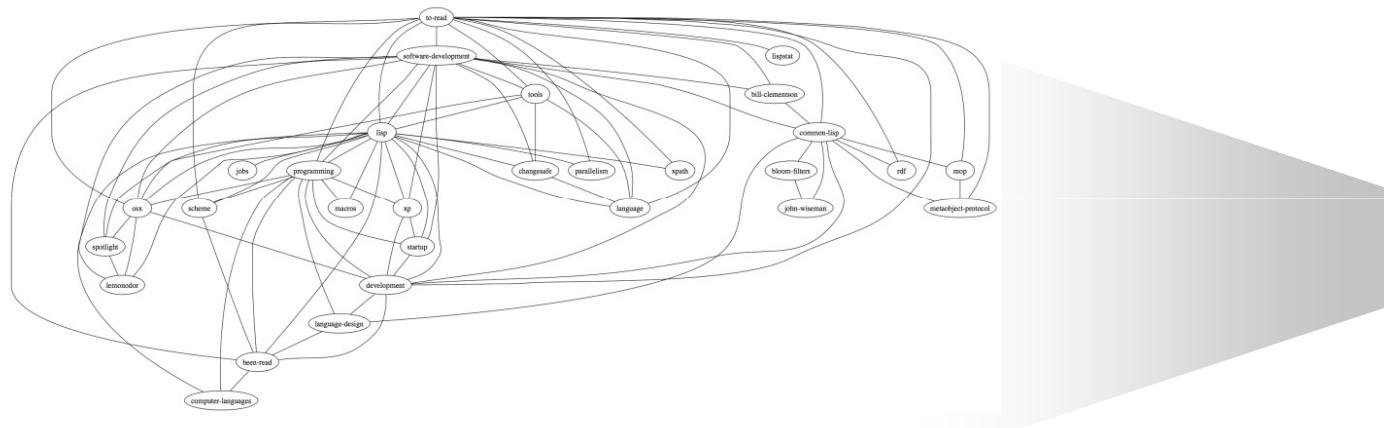
# The Richer, The Better

- Highly-interconnected semantic networks have a great impact on knowledge-based WSD even in a fine-grained setting [Navigli & Lapata, IEEE TPAMI 2010]



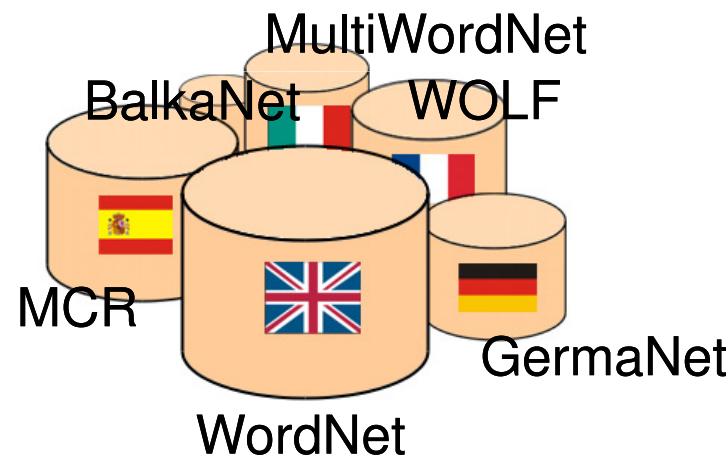
# Knowledge-based WSD NEEDS (a lot of) Knowledge!

- Knowledge-based approaches have a **high potential**
  - Lexical knowledge resources **only partly available**



# State of the Art “in a nutshell”

- Knowledge-based approaches have a **higher potential**
  - Lexical knowledge resources **only partly available**
  - Only for **few languages** (e.g. not all 23 EU official languages)
  - **Heterogenous** and with **low coverage**



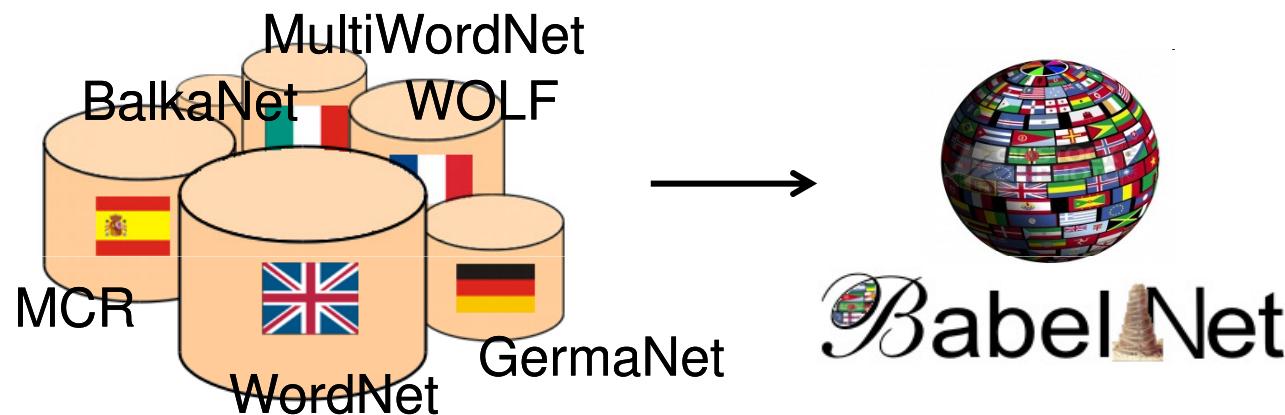
**This is where the ERC (and my project) comes into play**



A **5-year ERC Starting Grant (2011-2016)**  
on Multilingual Word Sense Disambiguation  
(<http://lcl.uniroma1.it/multijedi>)

# Multilingual Joint Word Sense Disambiguation (MultiJEDI)

**Key Objective 1:** create knowledge for all languages



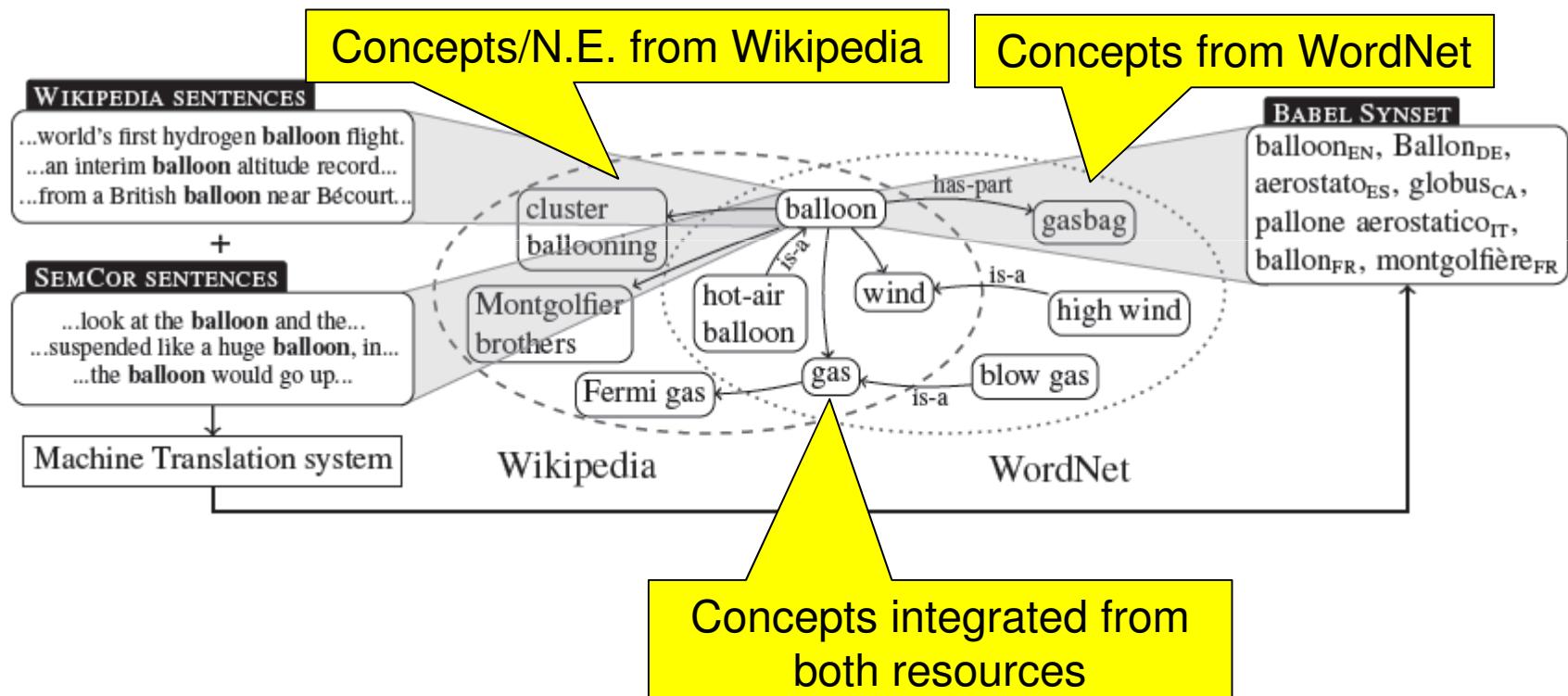
# Multilingual Joint Word Sense Disambiguation (MultiJEDI)

**Key Objective 2:** use all languages to disambiguate one



# BabelNet [Navigli & Ponzetto, ACL 2010; AIJ 2012]

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



# BabelNet integrates the best of both worlds

WordNet

balloon

Speech balloon

From Wikipedia, the free encyclopedia

Speech balloons (also speech bubbles, dialogue balloons or word balloons) are a graphic convention used most commonly in comic books, comic strips and cartoons to allow words (and much less often, pictures) to be understood as representing the speech or thoughts of a given character in the comic. There is often a formal distinction between the balloon that indicates thoughts and the one that indicates words spoken aloud: the balloon that conveys subjective thoughts is often as a thought balloon.



Balloon (typeface)

From Wikipedia, the free encyclopedia

This article does not cite any references or sources. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. (June 2012)

Contents [hide]

- 1 History
- 2 Popular forms
  - 2.1 Speech bubbles
  - 2.2 Thought bubbles
  - 2.3 Other forms

Balloon is a brush script commonly used for signage or display purposes. It was designed in 1939 by Max R. Kaufmann, for American Type Founders, in response to Howard Allen Trafton's Cartoon, cut for Bauer Type Foundry in 1936. It had no lowercase letters and was cast in Light, Bold, and Extra

**BALLOON**

Category Script

Designer(s) Max R. Kaufmann

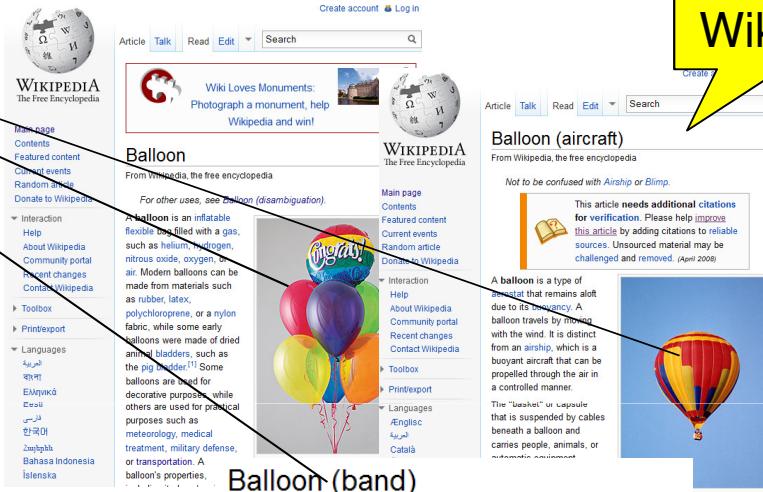
Foundry ATF

**THE QUICK BROWN FOX JUMPS OVER THE LAZY DOG.**

Sample

S: (n) balloon (large tough nonrigid bag filled with gas or heated air)  
S: (n) balloon (small thin inflatable rubber bag with narrow neck)

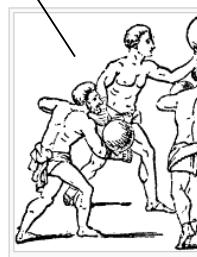
Wikipedia



Balloon (band)

From Wikipedia, the free encyclopedia

Balloon were an early 1990s duo from London, consisting of Ian Bickerton and David Sheppard. Their first and only album, *Gravity*, was released in 1992 by Dedicated, a British record label known for neo-psychadelia.<sup>[1]</sup> Produced by Michael Brook the record featured contributions from Sarah McLachlan (on the track "Tightrope Walker")<sup>[1]</sup> and James Pinker. Bickerton wrote the lyrics, while Sheppard provided the melodies and arrangements. The album was recorded mostly in New Orleans,<sup>[1]</sup> at Daniel Lanois' studio. The duo toured the US in 1992,<sup>[2]</sup> with percussionist James Pinker as a touring member.<sup>[3]</sup>



Balloon (game)

From Wikipedia, the free encyclopedia

Balloon, balloon-ball or wind-ball was a game similar to the modern game of volleyball in which a leather ball would be batted by the fist or forearm to prevent it from touching the ground. The game was played in ancient Rome where it was known as *follis* — the Latin word for a leather bag. Such a ball made of leather was quite heavy and so protection might be used such as a leather gauntlet or

21/09/2012

20

Babelplagiarism: What can BabelNet do for cross-language plagiarism detection?

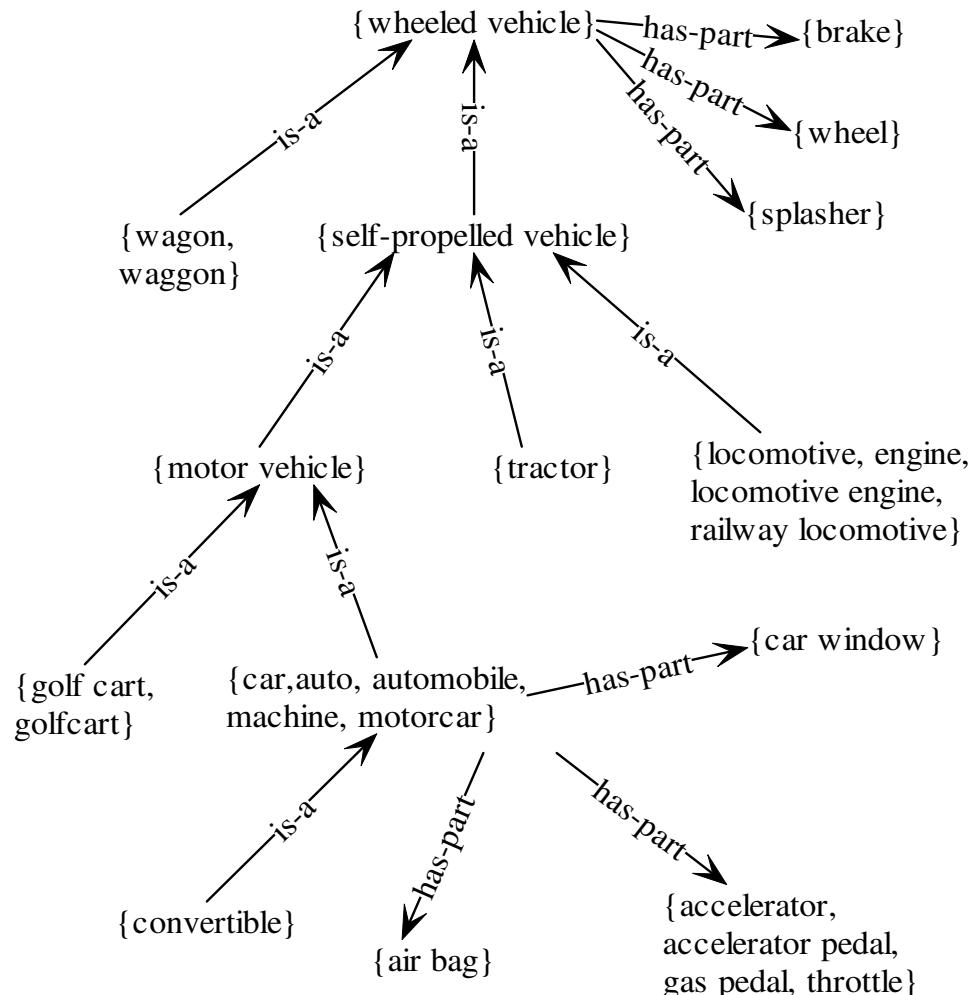
Roberto Navigli

# WordNet [Miller et al., 1990; Fellbaum, 1998]

## WordNet

- The most widespread computational lexicon of English [Miller et al., 1990]
- Concepts are encoded as set of synonyms (**synsets**), e.g.:  
 $\{ \text{pop}_n^2, \text{soda}_n^2, \text{soda pop}_n^1, \text{soda water}_n^2, \text{tonic}_n^2 \}$
- Semantic relations connect pairs of synsets
- For each synset, a textual definition (**gloss**) is provided, e.g.:  
“a sweet drink containing carbonated water and flavoring”.

# WordNet [Miller et al., 1990; Fellbaum, 1998]



# Wikipedia [the online community, 2001-today]

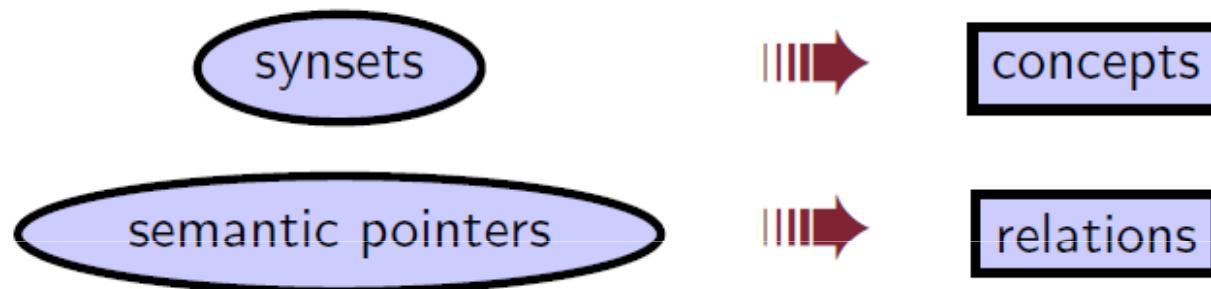
## Wikipedia

- The largest Web encyclopedia
- Wikipedia pages (*Wikipages*) encode: concepts (SODA (SOFT DRINK)) or named entities (FOOD STANDARDS AGENCY)
- The title of a Wikipage (e.g. SODA (SOFT DRINK)) is composed of:
  - lemma (soda)
  - possibly, a sense label (soft drink vs. sodium carbonate)
- Wikipages contain hyperlinks to other Wikipages
- Some Wikipages are redirections to other pages (e.g. SODA (SODIUM CARBONATE) → SODIUM CARBONATE)
- Wikipages are manually categorized (e.g. SOFT DRINKS for SODA)

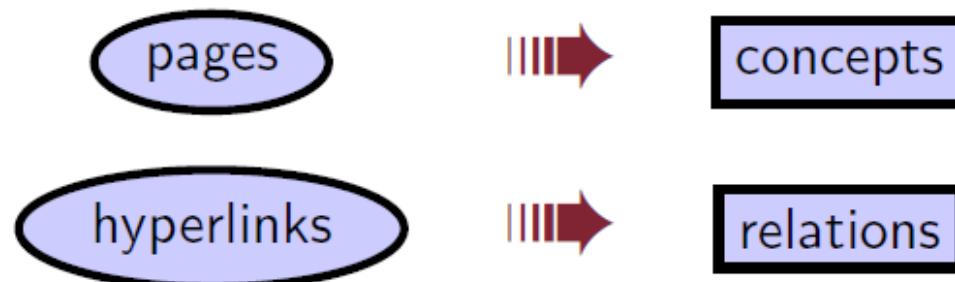
# BabelNet: concepts and semantic relations (1)

- Concepts and relations in BabelNet are harvested from **WordNet** and **Wikipedia**:

- **WordNet**



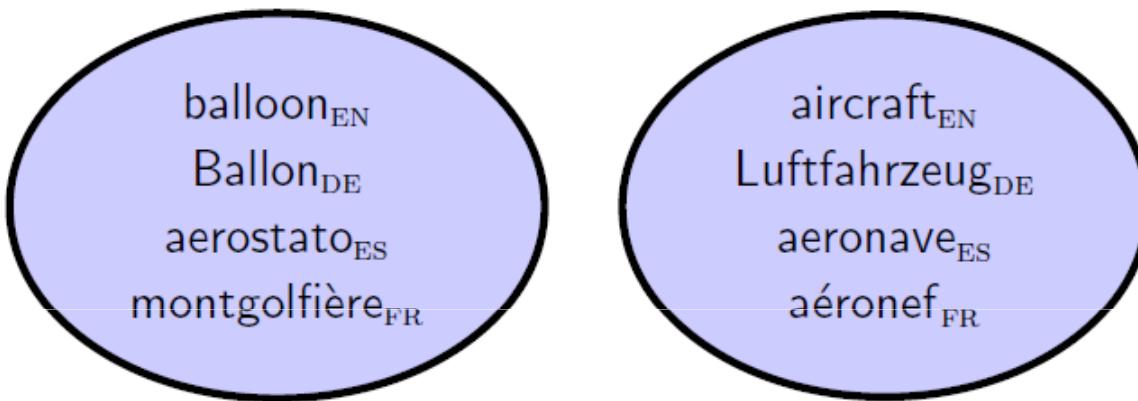
- **Wikipedia**



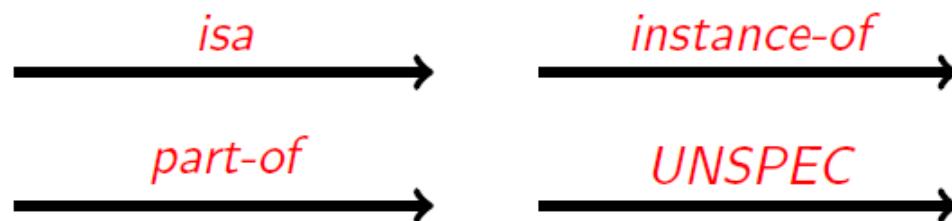
## BabelNet: concepts and semantic relations (2)

We encode knowledge as a labeled directed graph

- each vertex represents a **Babel synset**



- each edge expresses a **semantic relation**



# BabelNet: objectives

## 1. Provide a unified resource

- By establishing an automated mapping between Wikipedia pages and WordNet senses

## 2. Enable multilinguality

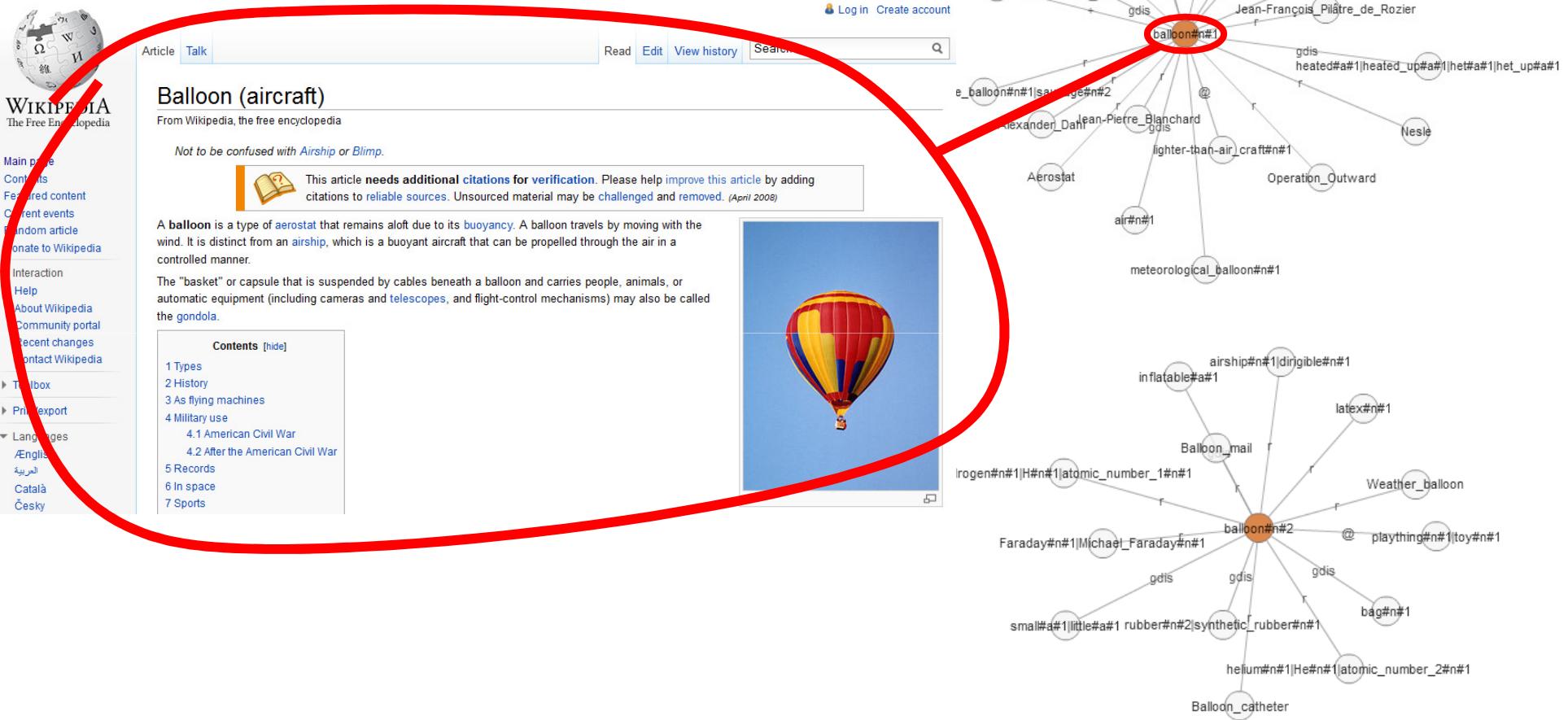
- By collecting the lexicalizations of concepts in different languages using:
  - a) Wikipedia interlanguage links
  - b) Statistical Machine Translation

# Building BabelNet: Mapping Wikipedia to WordNet (1)

- Bunescu & Pasca [2006] and Mihalcea [2007] used Wikipedia pages as **word senses**
- Mihalcea [2007] **manually mapped** Wikipedia pages to WordNet senses and performs lexical-sample WSD
- **Our contribution:** we fully **automatize** the mapping between Wikipedia and WordNet
  - We select the most likely WordNet sense  $s$  of a wikipedia page  $w$ :

$$\mu(w) = \begin{cases} s \in Senses_{WN}(w) & \text{if a link can be established,} \\ \epsilon & \text{otherwise.} \end{cases}$$

# An example of mapping



# Creation of the Wikipedia disambiguation contexts

- Wikipedia: given a page (e.g. BALLOON (AIRCRAFT))
  - ➡ sense labels aircraft
  - ➡ links wind, gas, helium, ...
  - ➡ categories technology

## Building BabelNet: Mapping Wikipedia to WordNet (2)

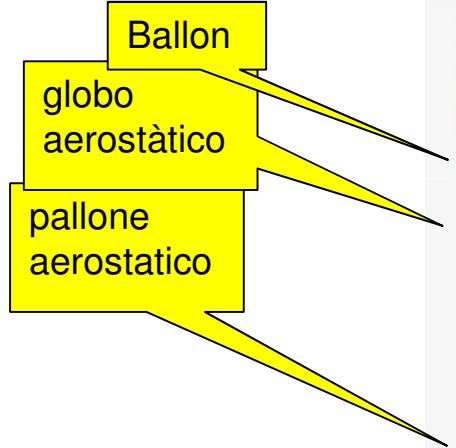
$$\begin{aligned}\mu(w) &= \operatorname{argmax}_{s \in Senses_{WN}(w)} p(s|w) = \operatorname{argmax}_s \frac{p(s, w)}{p(w)} \\ &= \operatorname{argmax}_s p(s, w)\end{aligned}$$

- Given a Wikipage  $w$  and its disambiguation context  $\text{ctx}(w)$ :
  - For each WordNet sense  $s$  of  $w$ , calculate  $\text{score}(s, w)$  as follows:

$$\text{score}(s, w) = \sum_{cw \in \text{Ctx}(w)} \sum_{s' \in Senses_{WN}(cw)} \sum_{p \in paths_{WN}(s, s')} e^{-(length(p)-1)}$$

# Building BabelNet: Translating Babel synsets

## 1. Exploiting Wikipedia interlanguage links



automatic equipment (including cameras and telescopes, and flight-control mechanisms) may also be called the gondola.

**Contents [hide]**

- 1 Types
- 2 History
- 3 As flying machines
- 4 Military use
  - 4.1 American Civil War
  - 4.2 After the American Civil War
- 5 Records
- 6 In space
- 7 Sports
- 8 See also
- 9 References
- 10 External links

**Types** [edit]

There are three main types of balloons:

- hot air balloons obtain their buoyancy by heating the air inside the balloon. They are the most common type of balloon aircraft. "Hot air balloon" is sometimes used incorrectly to denote any balloon that carries people.
- gas balloons are inflated with a gas of lower molecular weight than the ambient atmosphere. Most gas balloons operate with the internal pressure of the gas the same as the pressure of the surrounding atmosphere. There is a type of gas balloon, called a superpressure balloon, that can operate with the lifting gas at pressure that exceeds the pressure of the surrounding air, with the objective of limiting or eliminating the loss of gas from day-time heating. Gas balloons are filled with gases such as:
  - hydrogen – not widely used for aircraft since the Hindenburg disaster because of high flammability (except for some sport balloons as well as nearly all unmanned scientific and weather balloons).
  - helium – the gas used today for all airships and most manned balloons.
  - ammonia – used infrequently due to its caustic qualities and limited lift.
  - coal gas – used in the early days of ballooning; it is highly flammable.
  - methane – used as a lower cost lifting gas, but offering less lift than helium or hydrogen.<sup>[1]</sup>
- Rozière balloons use both heated and unheated lifting gases. The most common modern use of this type of balloon is for long-distance record flights such as the recent circumnavigations.

**History** [edit]

Main article: *History of ballooning*



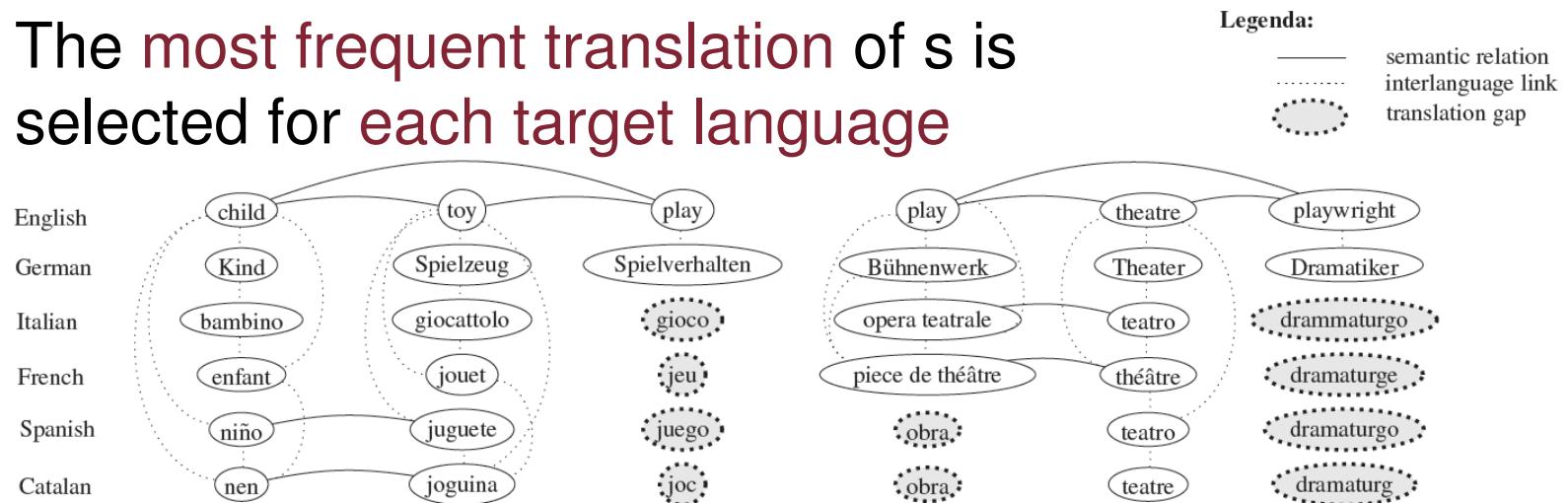
## Building BabelNet: Translating Babel synsets

2. Filling the lexical translation gaps using a Machine Translation system to translate the English lexicalizations of a concept
- On August 27, 1783 in Paris, Franklin witnessed the world's first hydrogen [[**Balloon (aircraft)**|**balloon**]] flight.
  - Le 27 Août, 1783 à Paris, Franklin vu le premier vol en **ballon** d'hydrogène.
- 

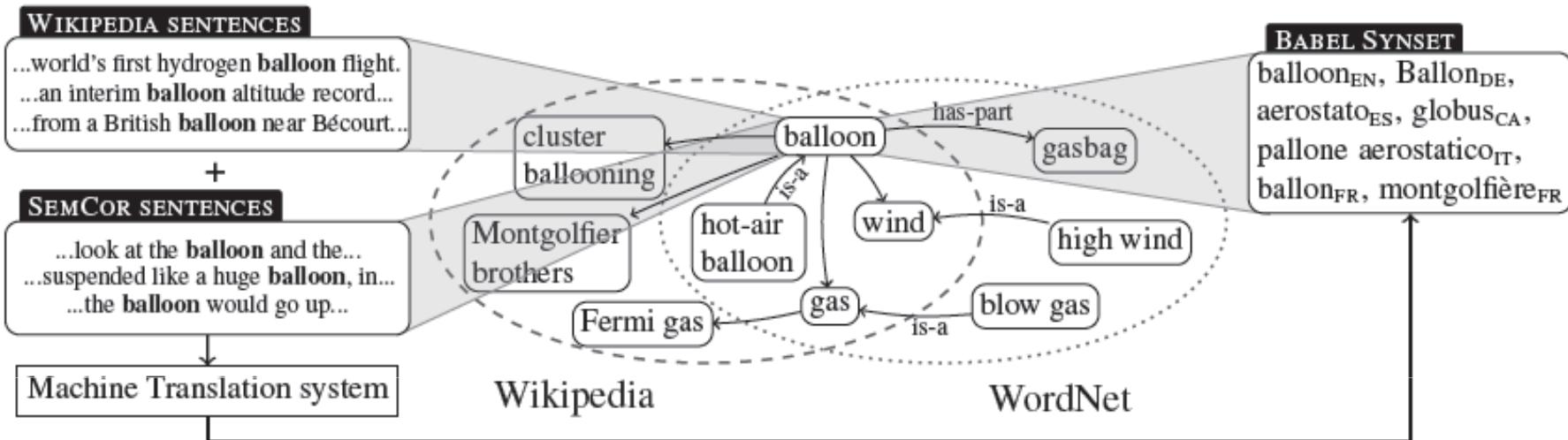
# Building BabelNet: Translating Babel synsets

## 2. Filling the lexical translation gaps using a Machine Translation system to translate the English lexicalizations of a concept

- For each word sense  $s$ , we translate:
  - sentences from **SemCor** (a corpus annotated with WordNet senses) which contain  $s$
  - sentences from **Wikipedia** linked to the Wikipage of  $s$
- The most frequent translation of  $s$  is selected for each target language



# BabelNet: an encyclopedic dictionary!



- Available online: <http://babelnet.org>

For research purposes...



# Anatomy of BabelNet

- 6 languages covered (moving to 40+)
- More than 3 million Babel synsets (i.e. concepts and NE)
- More than 26 million word senses:

	English	Catalan	French	German	Italian	Spanish	Total
English WordNet	206,978	-	-	-	-	-	206,978
Wikipedia	2,955,552	123,101	524,897	506,892	404,153	349,375	4,863,970
	3,388,049	105,147	617,379	456,977	217,963	404,009	5,189,524
WordNet	-	3,445,470	2,844,751	2,841,916	3,046,325	3,083,427	15,261,889
	-	97,876	98,081	97,672	98,475	98,092	490,196
SemCor	-	6,852	6,855	6,850	6,856	6,855	34,268
Total	6,550,579	3,778,446	4,091,963	3,910,307	3,773,772	3,941,758	26,046,825

- About 70 million lexico-semantic relations:

	English	Catalan	French	German	Italian	Spanish	Total
WordNet	364,552	-	-	-	-	-	364,552
WordNet glosses	617,785	-	-	-	-	-	617,785
Wikipedia	50,104,898	971,379	5,594,590	5,931,099	3,598,733	3,397,754	69,598,453
Total	51,087,235	971,379	5,594,590	5,931,099	3,598,733	3,397,754	70,580,790

# Evaluation of the Wikipedia-WordNet mapping

- Test set of 1,000 Wikipages manually mapped to the corresponding WordNet sense, if available

Mapping method		P	R	F <sub>1</sub>	A	
BoW	taxonomic	<b>89.7</b>	47.8	62.3	72.6	
	gloss	87.6	51.8	65.1	74.0	
	taxonomic + gloss	87.5	65.6	75.0	80.9	
Graph	taxonomic relations					
	max depth {@ 2 @ 3 @ 4}	87.2	60.8	71.6	77.9	
		81.6	65.0	72.4	78.7	
		<u>78.3</u>	<u>69.5</u>	<u>73.6</u>	<u>79.4</u>	
	gloss relations					
	max depth {@ 2 @ 3 @ 4}	80.5	60.6	69.1	77.0	
		<u>77.5</u>	<u>65.2</u>	<u>70.9</u>	<u>78.2</u>	
		72.4	67.1	69.6	78.0	
	taxonomic + gloss relations					
	max depth {@ 2 @ 3 @ 4}	81.2	74.6	<b>77.7</b>	<b>82.7</b>	
		72.8	<b>77.4</b>	75.1	80.1	
		64.3	76.2	69.8	75.0	
		MFS baseline	25.4	49.2	33.5	
		Random baseline	24.2	46.9	31.9	
					24.2	

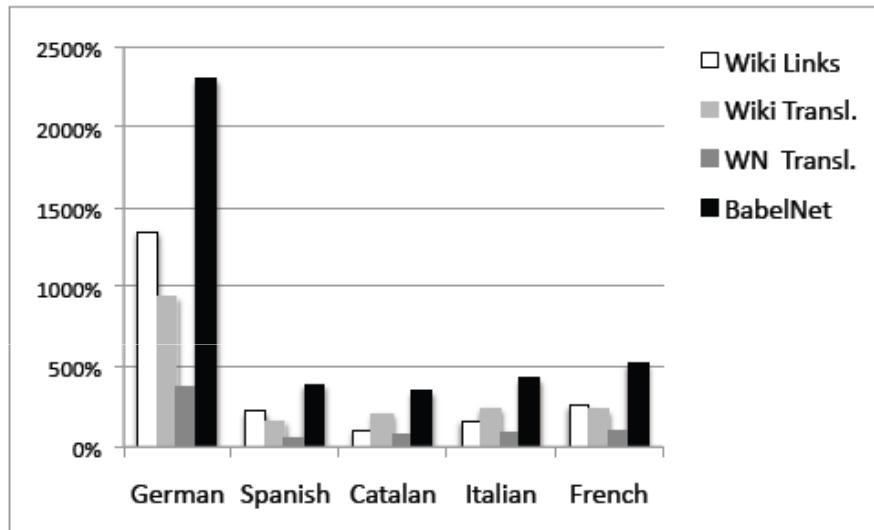
# Evaluation of BabelNet against gold standard resources

## Coverage

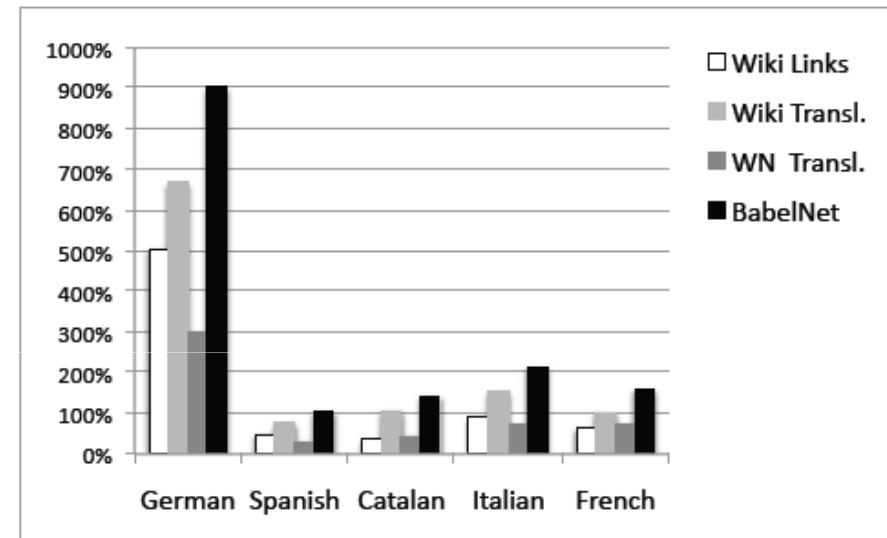
	Resource	Method	SENSES	SYNSETS
German	WIKI	{ Links Transl.	39.6 42.6	50.7 58.2
	WN	Transl.	21.0	28.6
	BABELNET	All	<b>57.6</b>	<b>73.4</b>
Spanish	WIKI	{ Links Transl.	34.4 47.9	40.7 56.1
	WN	Transl.	25.2	30.0
	BABELNET	All	<b>66.4</b>	<b>76.6</b>
Catalan	WIKI	{ Links Transl.	20.3 46.9	25.2 54.1
	WN	Transl.	25.0	29.6
	BABELNET	All	<b>64.0</b>	<b>73.3</b>
Italian	WIKI	{ Links Transl.	28.1 39.9	40.0 58.0
	WN	Transl.	19.7	28.7
	BABELNET	All	<b>52.9</b>	<b>73.7</b>
French	WIKI	{ Links Transl.	70.0 69.6	72.4 79.6
	WN	Transl.	16.3	19.4
	BABELNET	All	<b>86.0</b>	<b>92.9</b>

# Evaluation of BabelNet against gold standard resources

## Extra-coverage



(a) word senses



(b) synsets

# Coarse-grained Word Sense Disambiguation with BabelNet

Resource	Algorithm	Nouns only P/R/F <sub>1</sub>	All words P/R/F <sub>1</sub>
WordNet	Degree	80.1	79.7
	PLength	80.2	79.8
	SProbability	79.8	79.3
	PageRank	79.9	79.4
BabelNet	Degree	<b>84.7</b>	<b>82.3</b>
	PLength	<b>85.4</b>	<b>82.7</b>
	SProbability	<b>84.6</b>	<b>82.1</b>
	PageRank	82.3	80.1
	SUSSX-FR	81.1	77.0
	TreeMatch	N/A	73.6
	NUS-PT	82.3	<b>82.5</b>
	SSI	<b>84.1</b>	<b>83.2</b>
	MFS BL	77.4	78.9
	Random BL	63.5	62.7

## Main alternatives to BabelNet

- **WikiNet** [Nastase et al., 2011]
  - a multilingual semantic network built from Wikipedia and including semantic relations between Wikipedia entities collected from the category network, infoboxes and article bodies
- **Universal WordNet** [de Melo & Weikum, 2009]
  - bootstrapped from WordNet and built by collecting evidence extracted from existing wordnets, translation dictionaries, and parallel corpora
- **MENTA** [de Melo & Weikum, 2010]
  - multilingual taxonomy containing 5.4 million entities, also built from WordNet and Wikipedia using a number of heuristics

Resource	Lemmas	Concepts	Word senses
UWN	822,212	117,659	1,595,763
MENTA { upper-level	837,627	82,115	845,210
full	–	5,379,832	–
WikiNet	11,721,594	3,707,718	14,200,945
BabelNet	23,936,234	3,032,406	26,045,741

## BabelNetXplorer: A Java API and a Visual Explorer [Navigli & Ponzetto, WWW 2012 DEMO]

- We developed the **BabelNet API** for effectively accessing multilingual semantic networks such as **BabelNet**
  - A **Java API** based on **Apache Lucene**
  - **Available at:** <http://babelnet.org>
- We created a **Web application** for visualizing and exploring semantic networks
  - Based on Cytoscape Web, a state-of-the-art visualization software
- **Available at:** <http://lcl.uniroma1.it/bnxplorer>

# The BabelNet API

```
BabelNet bn = BabelNet.getInstance();
System.out.println("SYNSETS WITH English word: \"bank\"");
List<BabelSynset> synsets = bn.getSynsets(Language.EN, "bank");
for (BabelSynset synset : synsets)
{
    System.out.print(" =>(" + synset.getId() + ") SOURCE: " + synset.getSource() +
                    "; WN SYNSET: " + synset.getWordNetOffsets() + ";\n" +
                    " MAIN LEMMA: " + synset.getMainLemma() + ";\n SENSES (German): { ");
    for (BabelSense sense : synset.getSenses(Language.DE))
        System.out.print(sense.toString() + " ");
    System.out.println("}\n -----");
    Map<IPointer, List<BabelSynset>> relatedSynsets = synset.getRelatedSynsets();
    for (IPointer relationType : relatedSynsets.keySet())
    {
        List<BabelSynset> relationSynsets = relatedSynsets.get(relationType);
        for (BabelSynset relationSynset : relationSynsets)
        {
            System.out.println("   EDGE " + relationType.getSymbol() +
                               " " + relationSynset.getId() +
                               " " + relationSynset.toString(Language.EN));
        }
    }
    System.out.println(" -----");
}
```

Retrieve all synsets with the English lemma “bank”

Print information about each synset

Get the (relation, synsets) map of the synset neighbours

Get the synsets related by a given relation type

Print the information of each related synset

# BabelNetXplorer: semantic network exploration

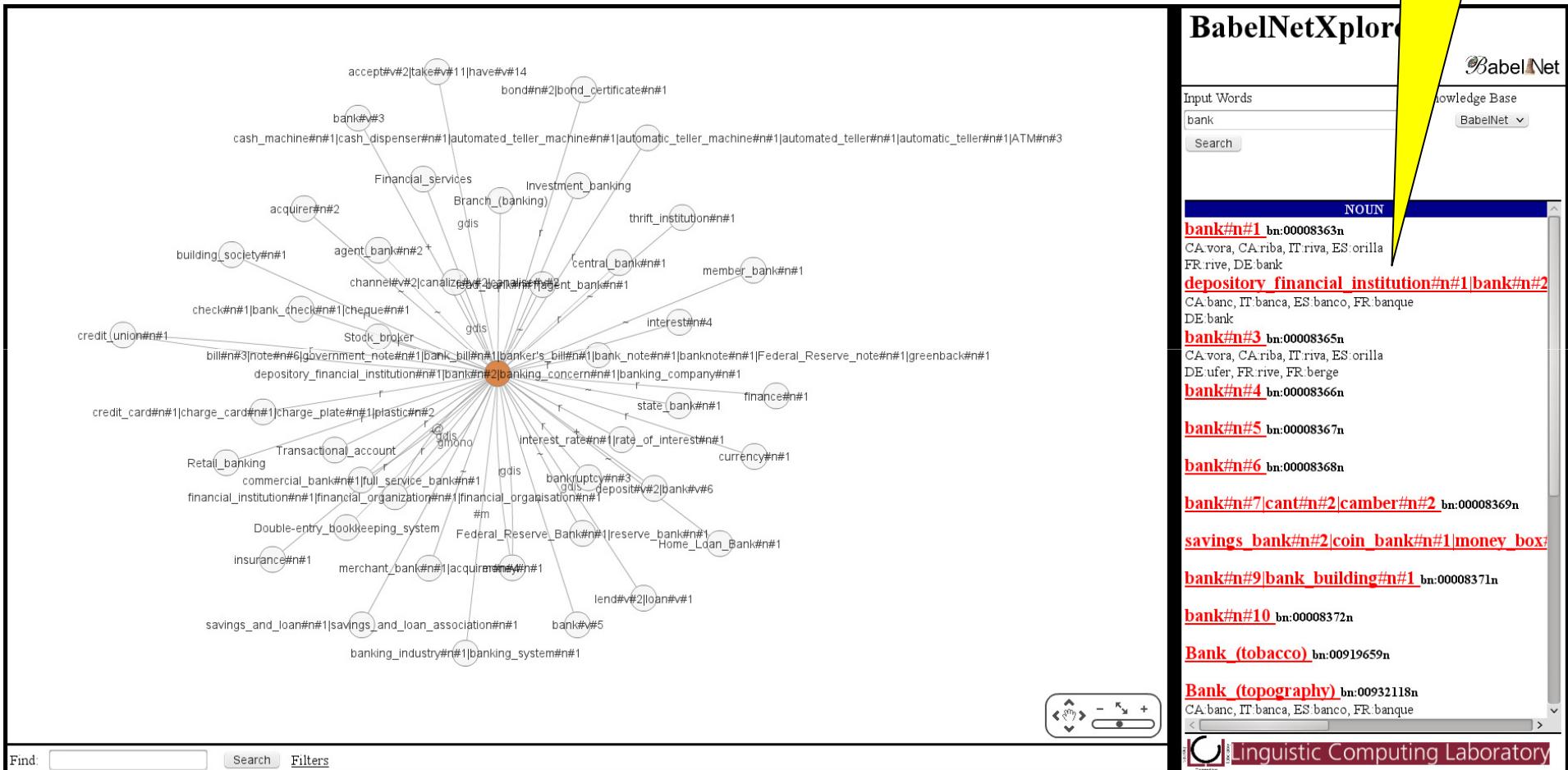
- Type a (possibly ambiguous) word in any language:

The screenshot shows the BabelNetXplorer interface. On the left, there is a large input field labeled "Input word" containing the word "bank". To the right, the results are displayed in a list format. The results are categorized by part of speech (NOUN). The first result is highlighted in red: "bank#n#1 bn:00008363n CA:vora, CA:riba, IT:riva, ES:orilla FR:rive, DE:bank". Below it, other results are listed in blue: "depository financial institution#n#1|bank#n#2 CA:banc, IT:banca, ES:banco, FR:banque DE:bank", "bank#n#3 bn:00008365n CA:vora, CA:riba, IT:riva, ES:orilla DE:ufer, FR:rive, FR:berge", "bank#n#4 bn:00008366n", "bank#n#5 bn:00008367n", "bank#n#6 bn:00008368n", "bank#n#7|cant#n#2|camber#n#2 bn:00008369n", "savings\_bank#n#2|coin\_bank#n#1|money\_box#n#1 bn:00008370n", "bank#n#9|bank\_building#n#1 bn:00008371n", "bank#n#10 bn:00008372n", "Bank\_(tobacco) bn:00919659n", "Bank\_(topography) bn:00932118n". At the bottom of the results, there is a logo for the "Linguistic Computing Laboratory" and a copyright notice "© 2012 Universitat de València".

# BabelNetXplorer: semantic network exploration

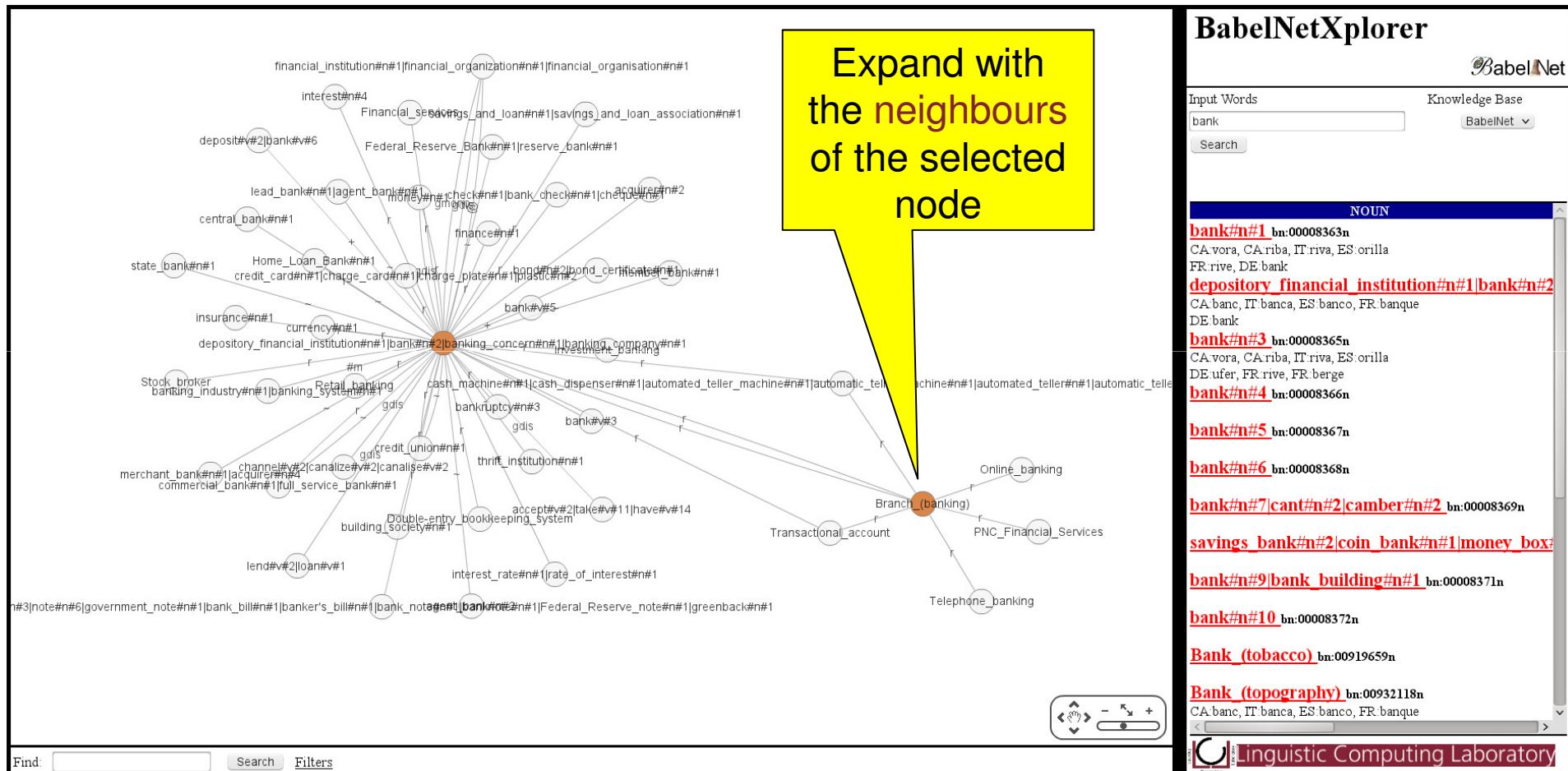
Selected sense

- Click a Babel sense of the input word:



# BabelNetXplorer: semantic network exploration

- **Expand the graph** by clicking on a node:



# Babelplagiarism: What can BabelNet do for cross-language plagiarism detection?

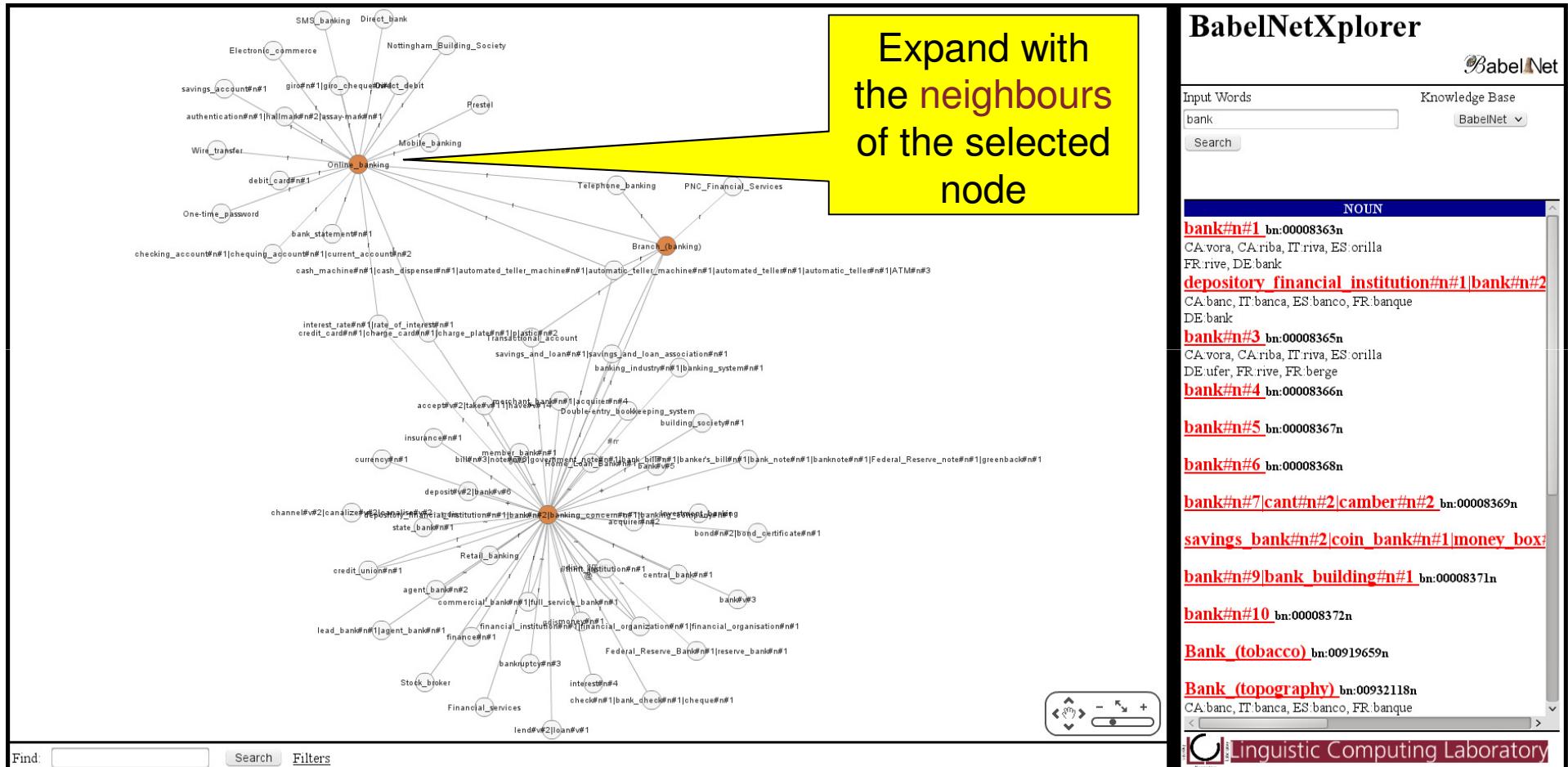
Roberto Navigli

21/09/2012

45

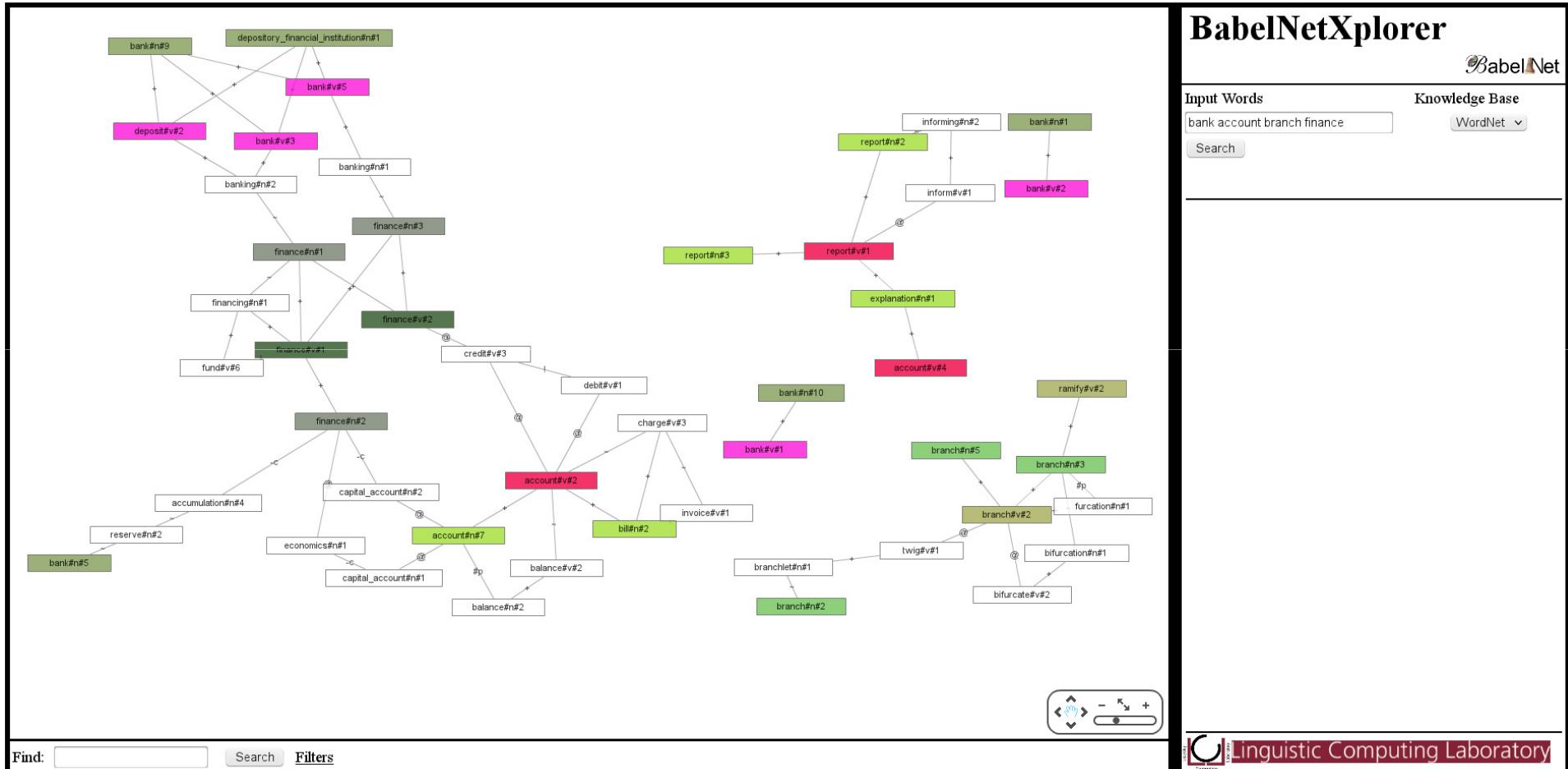
# BabelNetXplorer: semantic network exploration

- Expand the graph by clicking on a node:



# BabelNetXplorer: search for connecting paths

- Search the graph for connecting paths:



# Multilingual WSD with Just a Few Lines of Code

[Navigli & Ponzetto, ACL 2012 DEMO]

```
public static void acl12demo() throws IOException
{
    List<Word> sentence =
        Arrays.asList(new Word[] {
            new Word("bank", 'n', Language.EN),
            new Word("bonus", 'n', Language.EN),
            new Word("pay", 'v', Language.EN),
            new Word("stock", 'n', Language.EN)
        });
    disambiguate(sentence, KnowledgeBase.BABELNET, KnowledgeGraphScorer.DEGREE);
}

public static void disambiguate(Collection<Word> words,
                               KnowledgeBase kb,
                               KnowledgeGraphScorer scorer) throws IOException
{
    KnowledgeGraphFactory factory = KnowledgeGraphFactory.getInstance(kb);
    KnowledgeGraph kGraph = factory.getKnowledgeGraph(words);
    Map<String, Double> scores = scorer.score(kGraph);
    for (String concept : scores.keySet())
    {
        double score = scores.get(concept);
        for (Word word : kGraph.wordsForConcept(concept))
            word.addLabel(concept, score);
    }
    for (Word word : words)
    {
        System.out.println("\n\t" + word.getWord() + " -- ID " +
                           word.getId() + " => SENSE DISTRIBUTION: ");
        for (ScoredItem<String> label : word.getLabels())
        {
            System.out.println("\t [" + label.getItem() + "]:" +
                               Strings.format(label.getScore()));
        }
    }
}
```

Target words can even be  
in mixed languages!

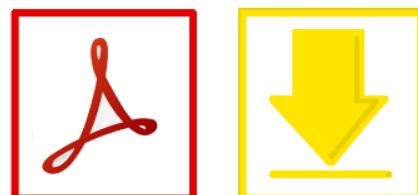
Create a disambiguation  
graph for the target words  
And disambiguate in 1 line!

# Coming soon to your screens: BabelNet 1.1!

**BabelNet**  
A very large multilingual ontology

means: 40 languages +  
more accurate mappings  
and translations!

search disambiguate



publications download



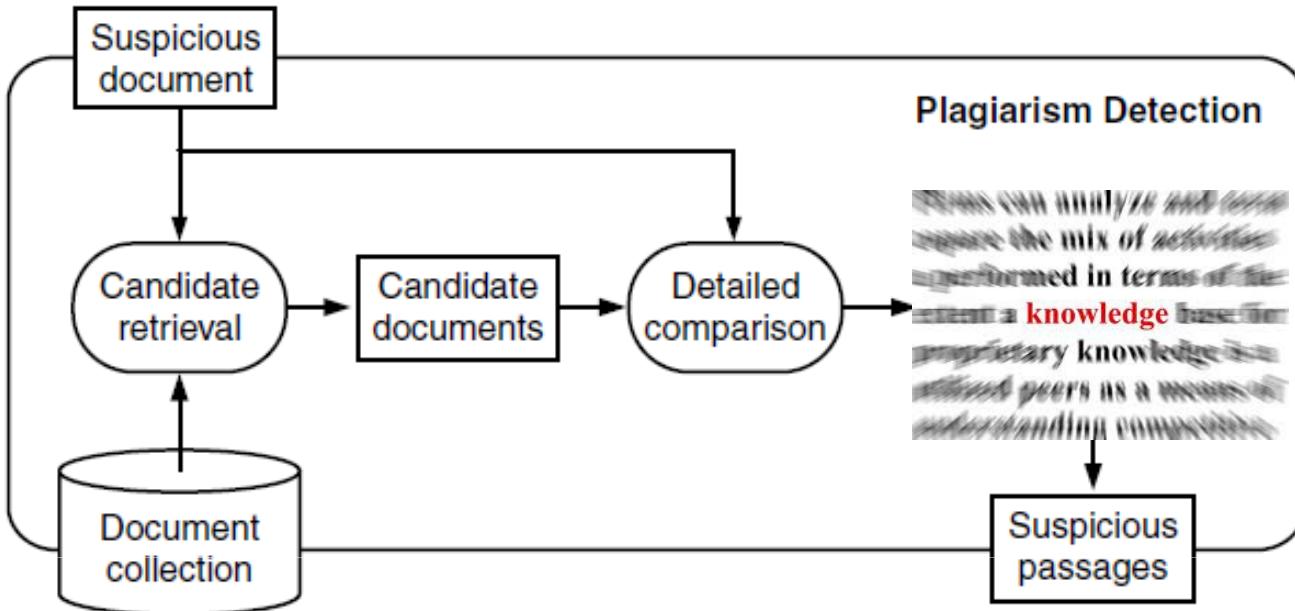
BabelNet is an output of the ERC Starting Grant MultiJEDI No. 259234.



# Now... why am I saying all this to YOU?!



# Plagiarism detection: the state of the art



[Stein et al., SIGIR 2007]

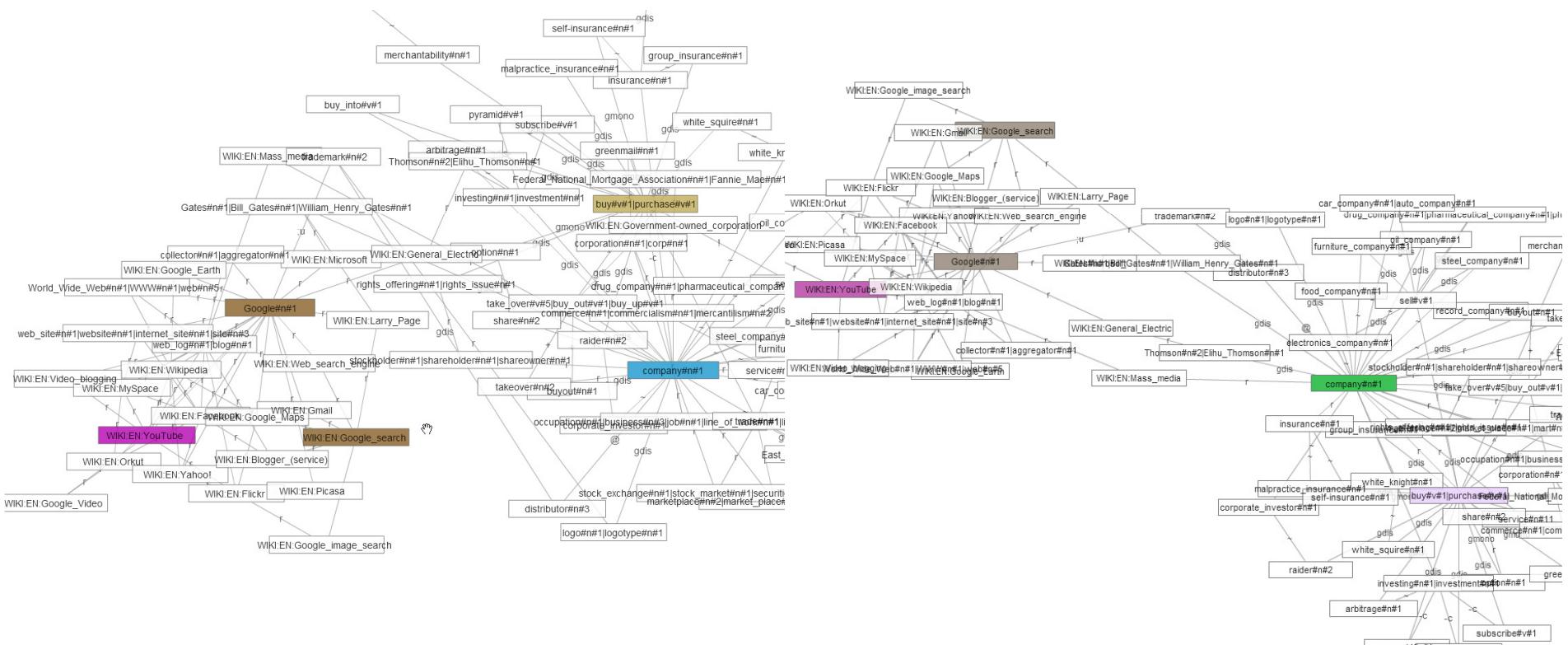
- Stemming, stopword removal, chunking into passages, keyphrase extraction, n-grams, query formulation, search control, etc.

**So, what can we do? [Examples from Vila et al. 2011]**

- **Same polarity substitutions:**

# Google bought YouTube

# Google purchased YouTube

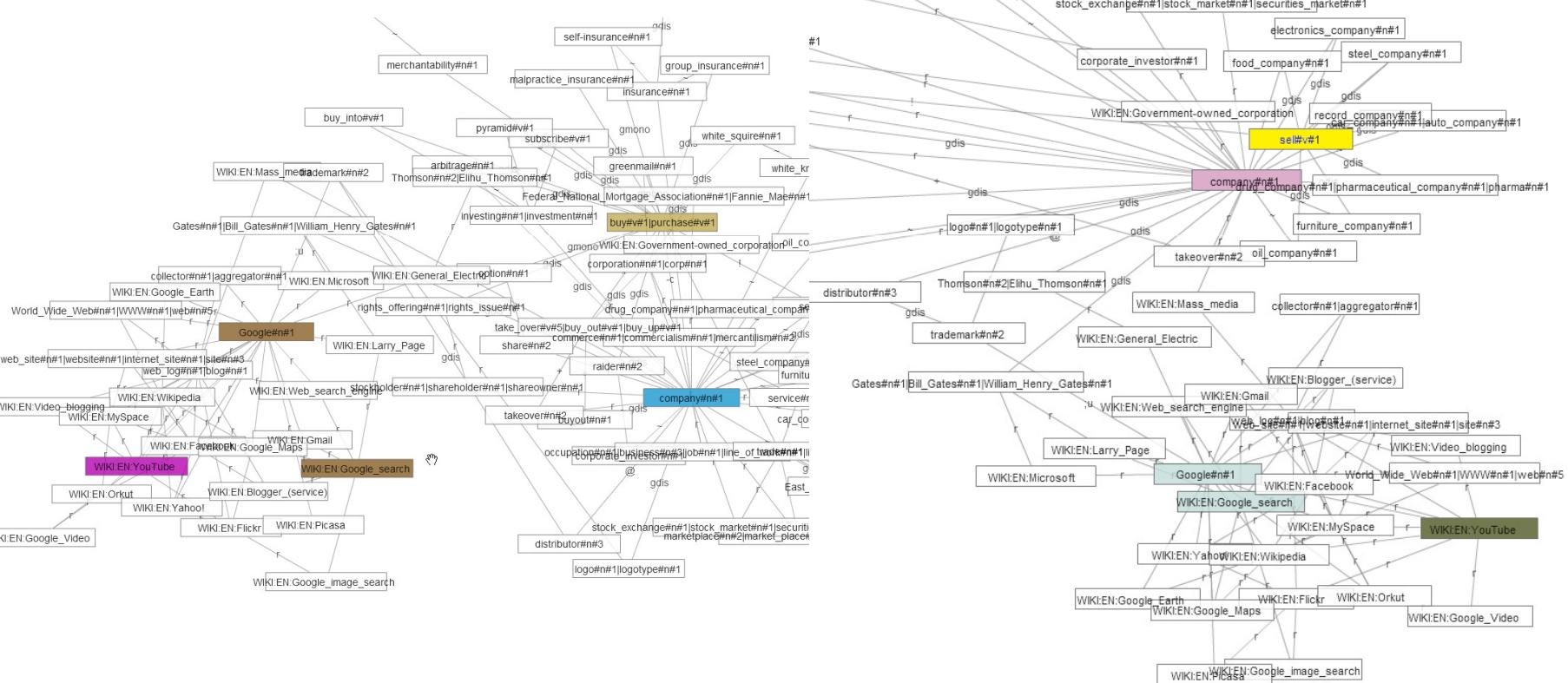


# So, what can we do?

- Opposite polarity substitutions:

Google bought YouTube

YouTube was sold to Google

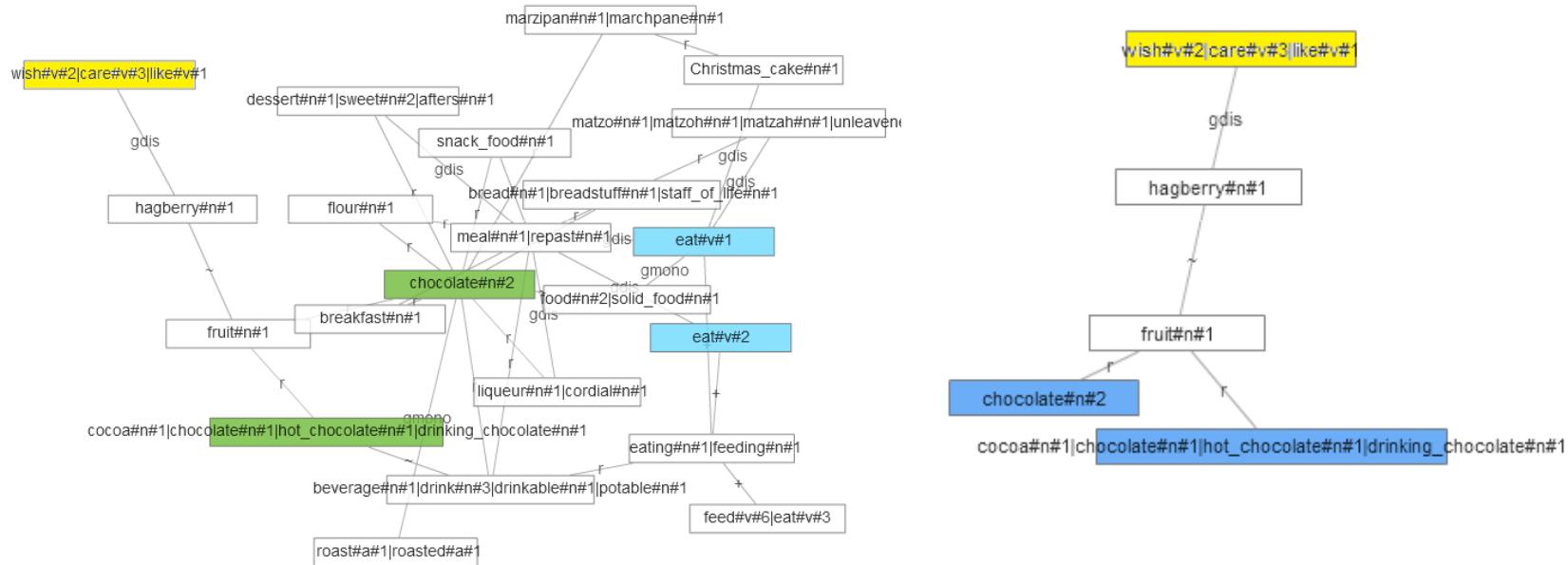


# So, what can we do?

- **Deletion:**

I like eating chocolate

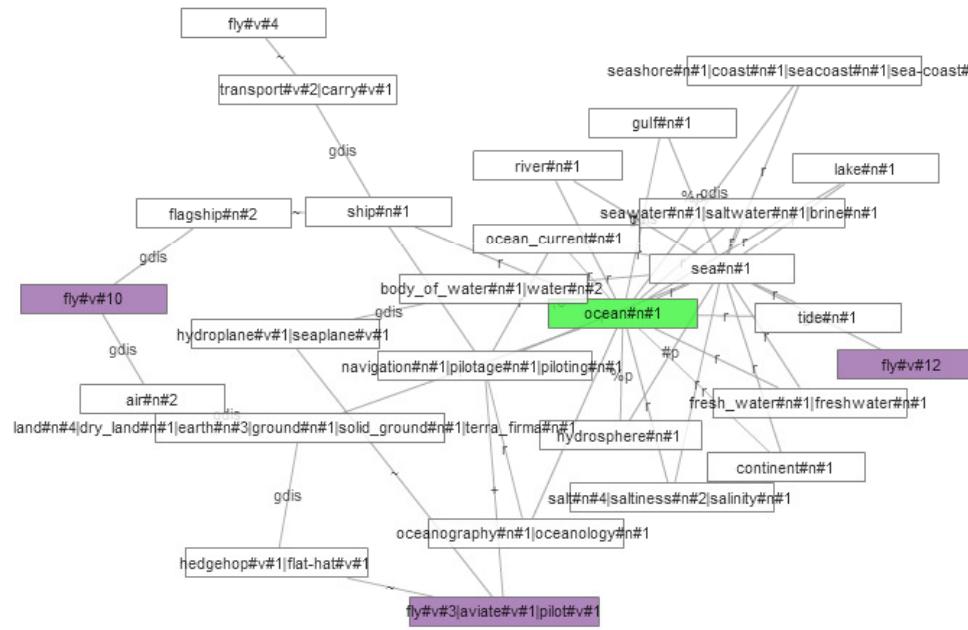
I like chocolate



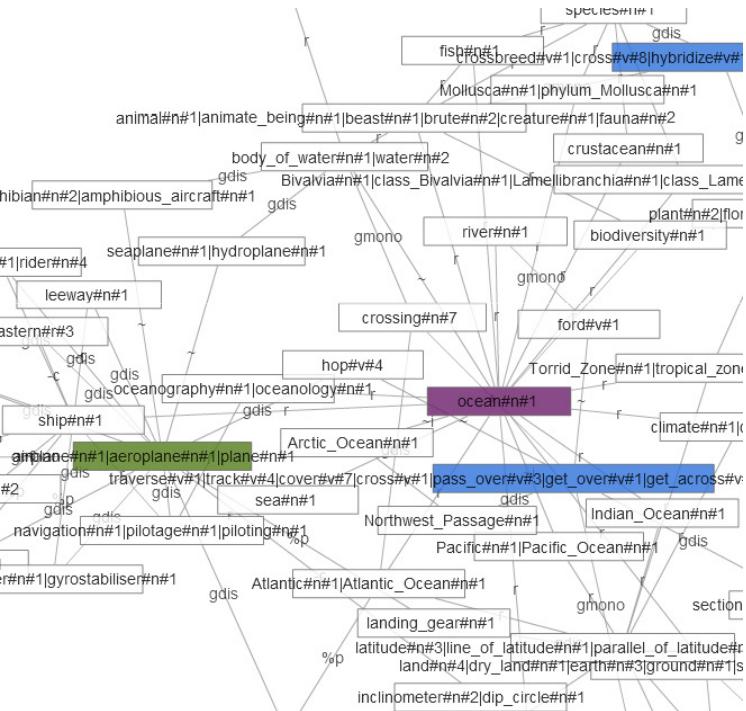
# So, what can we do?

- Semantics based changes:

Bill flew across the ocean



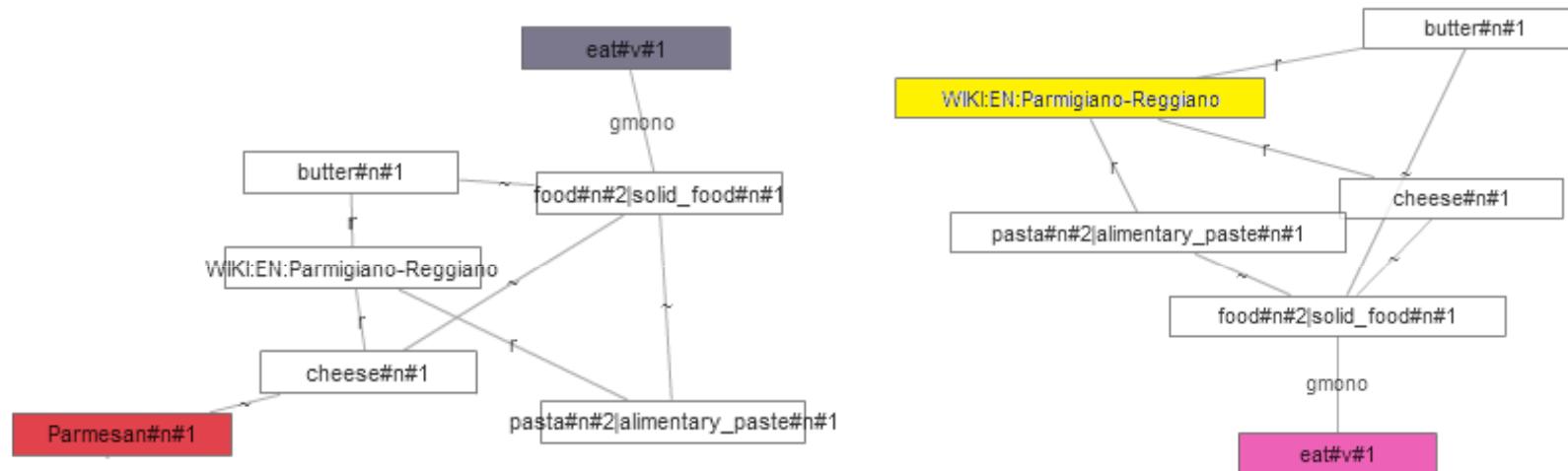
Bill crossed the ocean by plane



# Remember? BabelNet is multilingual!

- So one sentence can be in English, one in Italian

Paolo is eating Parmesan      Paolo sta mangiando il parmigiano



- However, note that only **nominal concepts and Named Entities are multilingual!**
  - verbs, adjectives and adverbs only in English

# Conclusions

- Statistics alone is not enough!
- We provide a (hopefully useful) tool for multilingual lexical semantics
- This includes cross-language plagiarism detection!
- You just have to download BabelNet and start coding!

```
unit cluedo;

interface
  procedure Solve;
implementation
  uses graderlib;
  procedure Solve;
  var
    x,y,z,t: longint;
  begin
    x:=1; y:=1; z:=1;
    t:=theory(x,y,z);
    while t <> 0 do
    begin
      if t = 1 then inc(x) else
      if t = 2 then inc(y)
      else inc(z);
      t:=theory(x,y,z);
    end;
  end;
begin
end.
```

# What comes next...



- Plenty of work to do!
- **BabelNet:**
  - Increasing the accuracy of BabelNet (e.g. game with a purpose)
  - Integrate more knowledge (Wikipedia categories, Wiktionary, adjectives, verbs, etc.)
  - Labeling relatedness relations (see WiSeNet [Moro & Navigli, CIKM 2012])
  - More languages (40+)
- Much more!

Thanks or...





SAPIENZA  
UNIVERSITÀ DI ROMA

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

Joint work with: Simone Ponzetto; +Mirella Lapata, +Andrea Moro

