

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

Roberto Navigli

DIPARTIMENTO
DI INFORMATICA

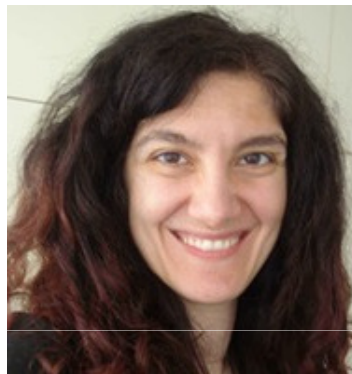


SAPIENZA
UNIVERSITÀ DI ROMA

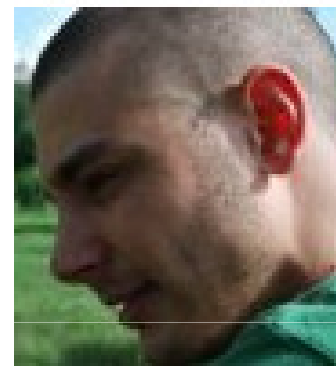
Joint work with...



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**?

...can analyze and even
represent the mix of activities
performed in terms of the
extent a **knowledge** base on
proprietary knowledge is a
valued peers as a means of
understanding competition

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.

Danger here!



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

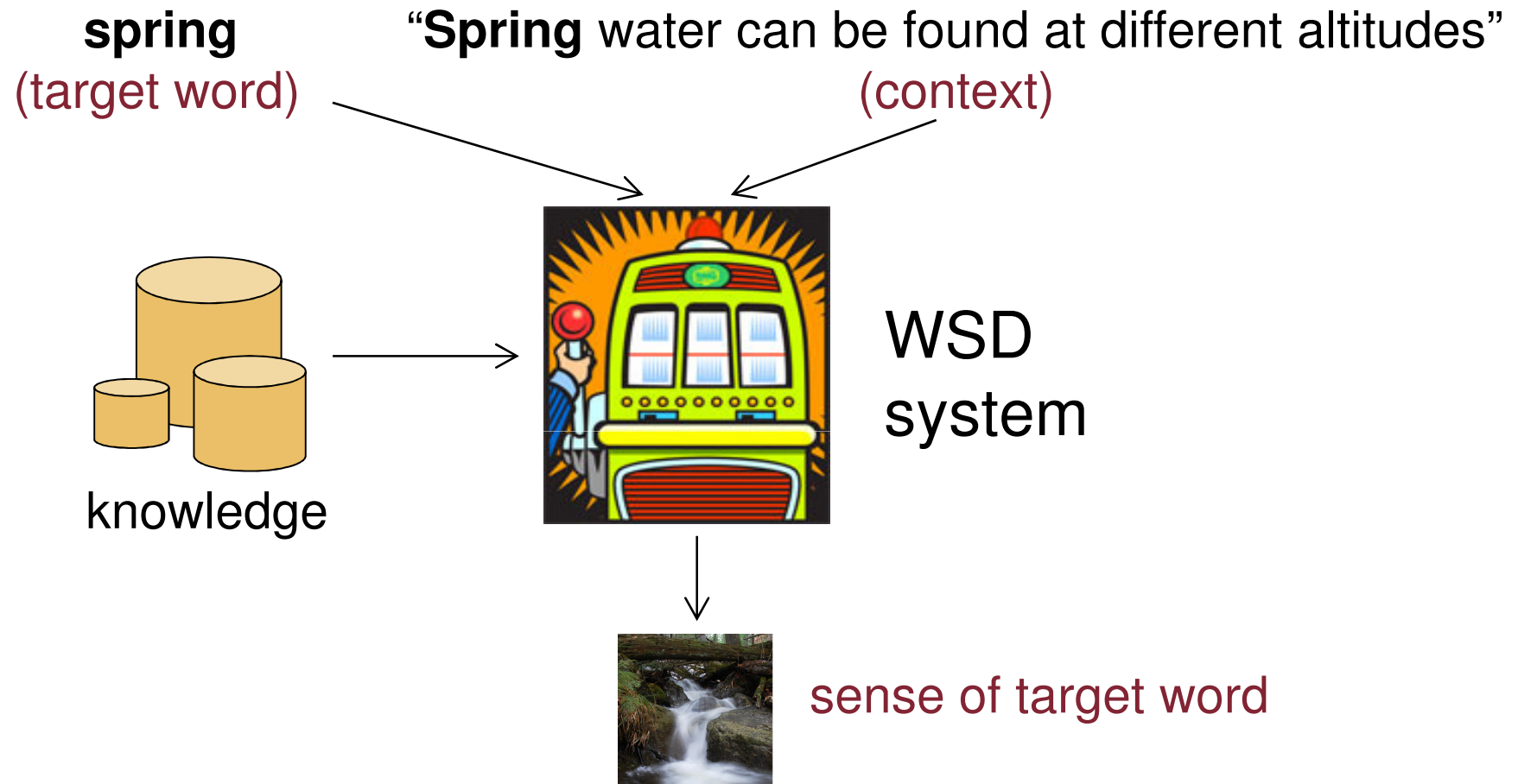


The Knowledge Acquisition Bottleneck

- **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!

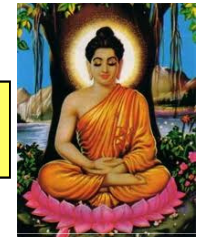
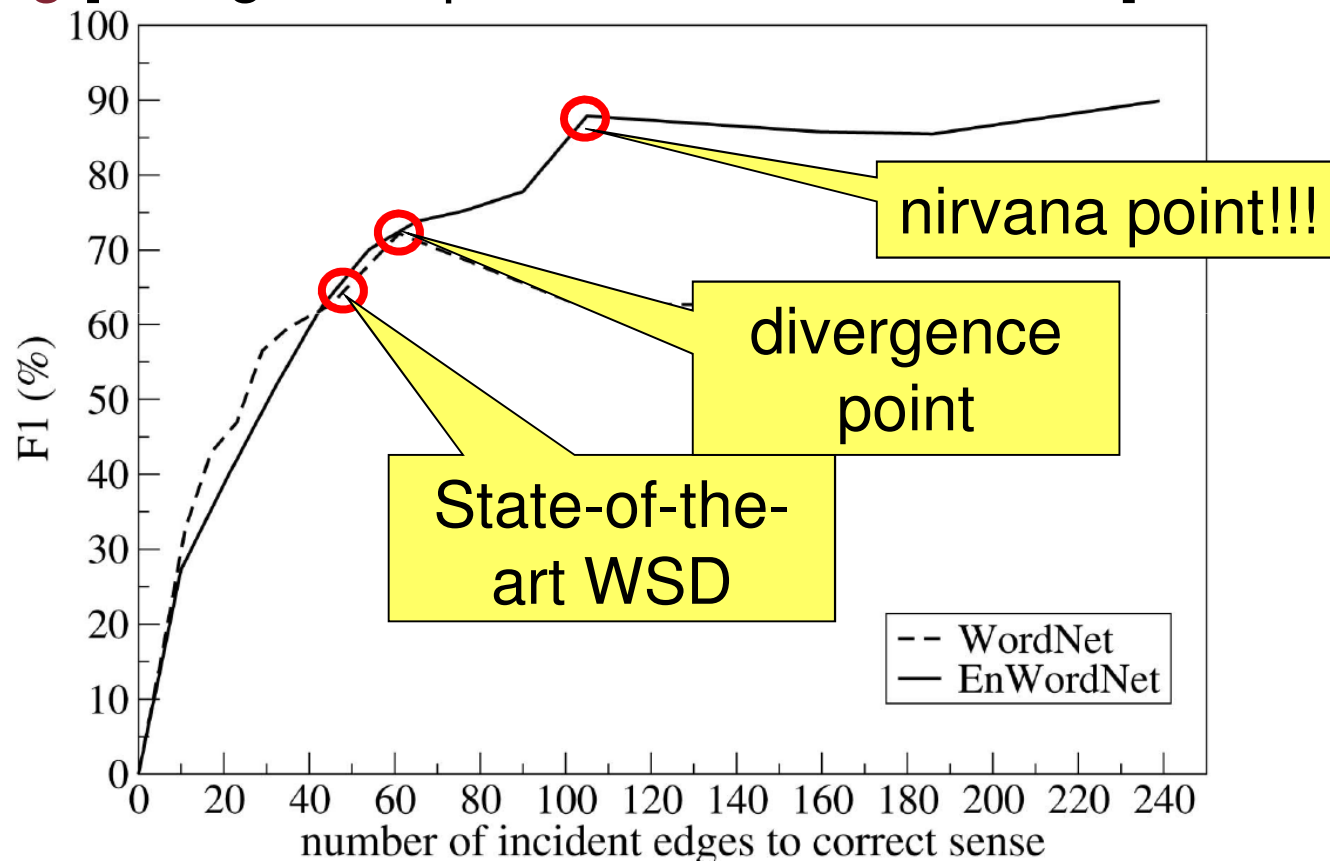
Word Sense Disambiguation in a Nutshell



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

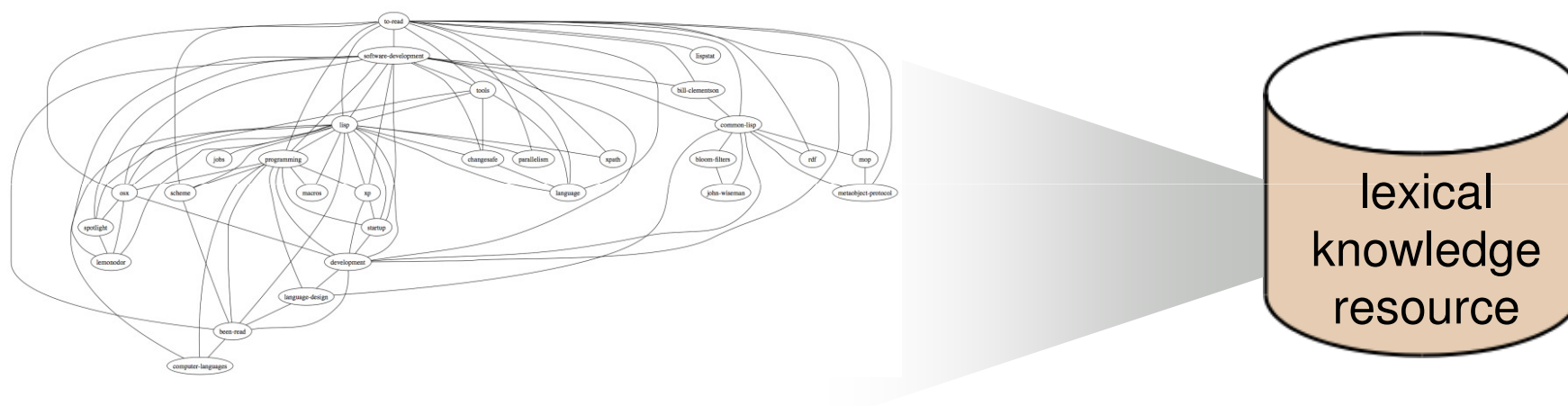
The Richer, The Better

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



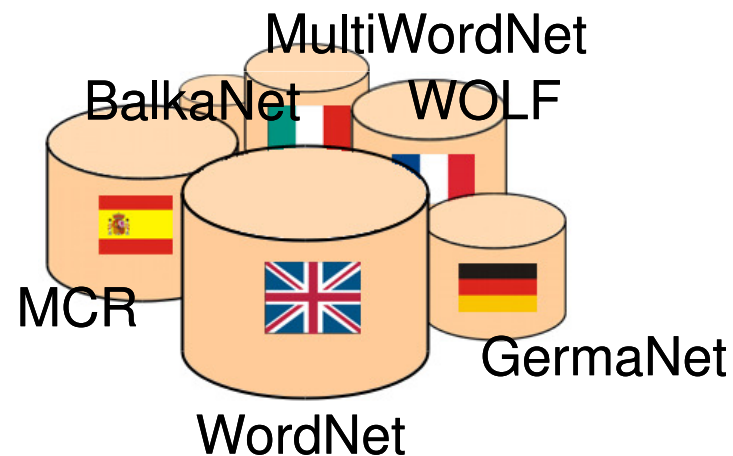
Knowledge-based WSD NEEDS (a lot of) Knowledge!

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

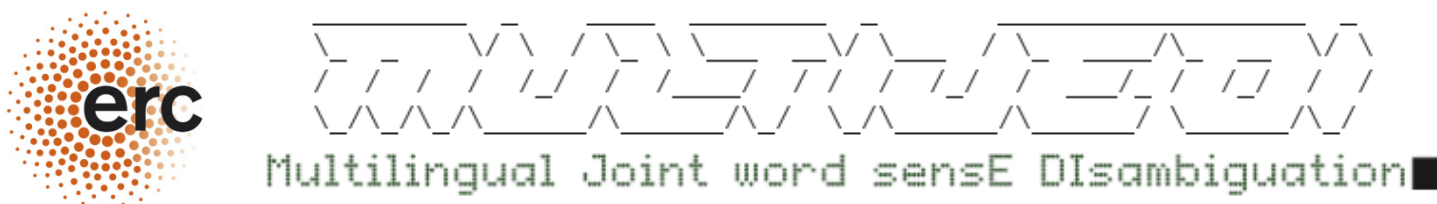


State of the Art “in a nutshell”

- 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**



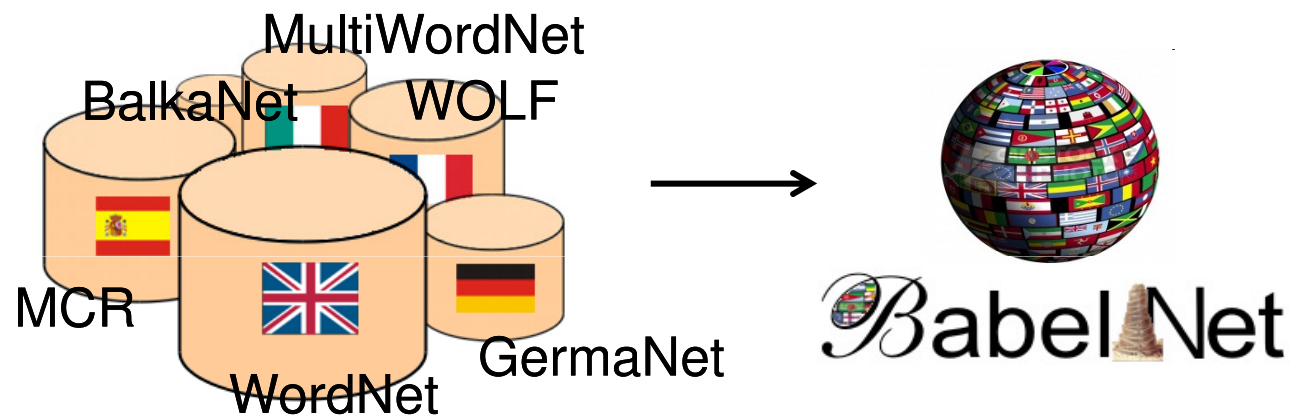
This is where the ERC (and my project) comes into play



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



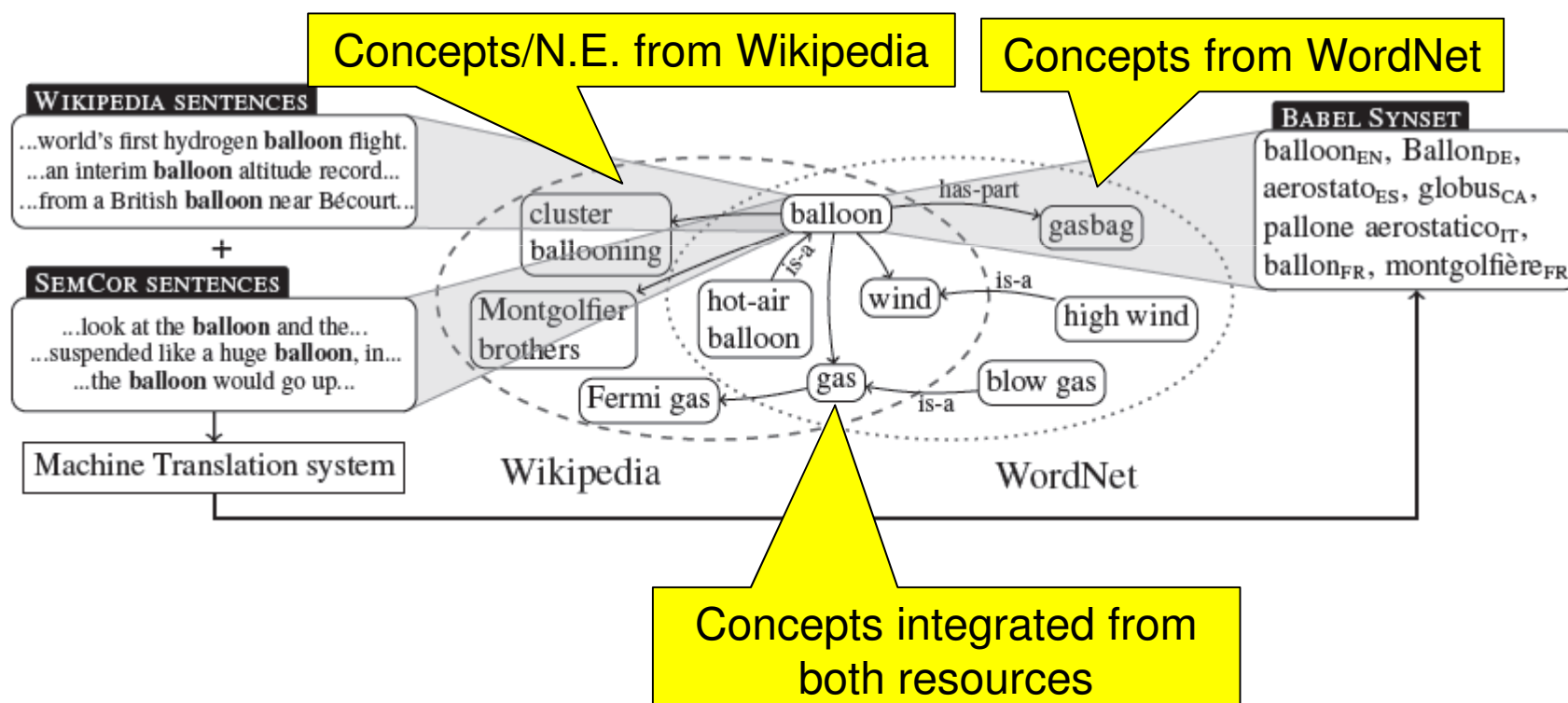
Multilingual Joint Word Sense Disambiguation (MultiJEDI)

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

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

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 latter 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)

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



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 *foliis* — 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



Ball game in progress

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-psychedelia**.^[1] Produced by **Michael Brook** the record featured contributions from **Sarah McLachlan** (on the track "Tightrope Walker")^[1] and James Pinder. Bickerton wrote the lyrics, while Sheppard provided the melodies and arrangements. The album was recorded mostly in **New Orleans**,^[1] at **Daniel Lanois'** studio. The duo toured the US in 1992,^[2] with percussionist James Pinder as a touring member.^[3]

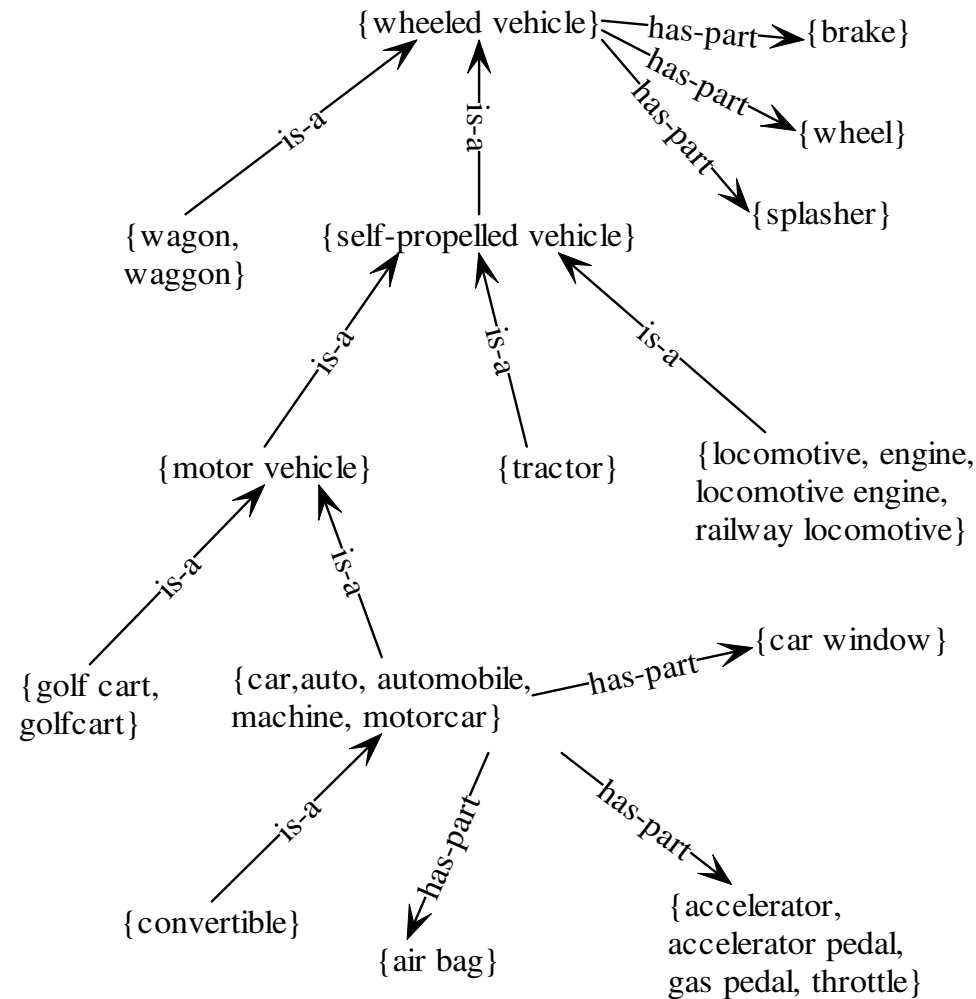
Bickerton and Sheppard met in 1988. They recorded more than 60 demo tracks before they were signed by Dedicated.^[2] Critic Eve Zibart of *The Washington Post* described the band's music as "a peculiarly soothing mix of Feargal Sharkey, Storyville, white soul, mild social unrest punk and Leonard Cohen."^[4] Jae-Ha Kim of the *Chicago Sun-Times* wrote, "Balloon's music is atmospheric and winsome, with acoustic guitars caressing velvety voices."^[5]

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)

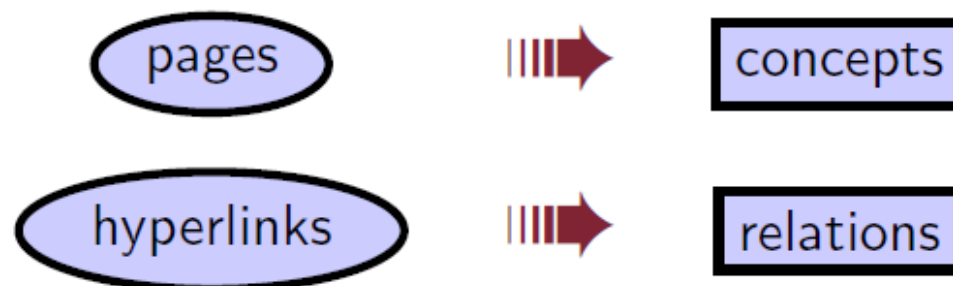
BabelNet: concepts and semantic relations (1)

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

- **WordNet**



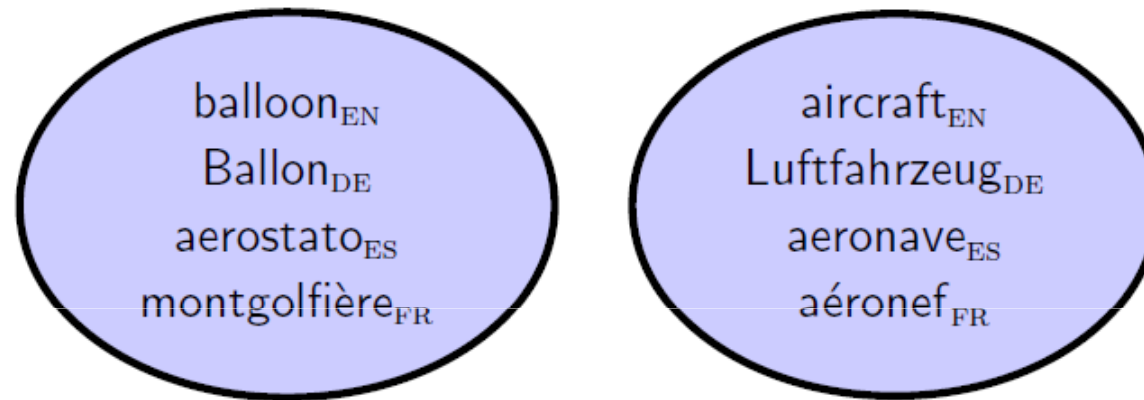
- **Wikipedia**



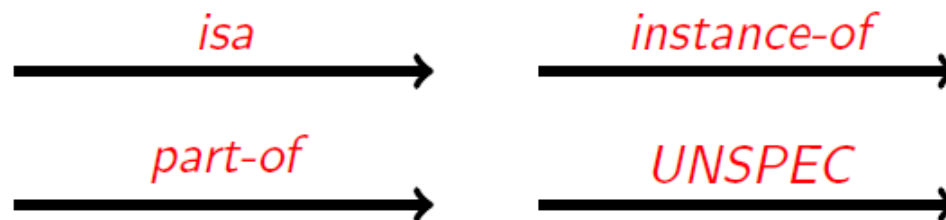
BabelNet: concepts and semantic relations (2)

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

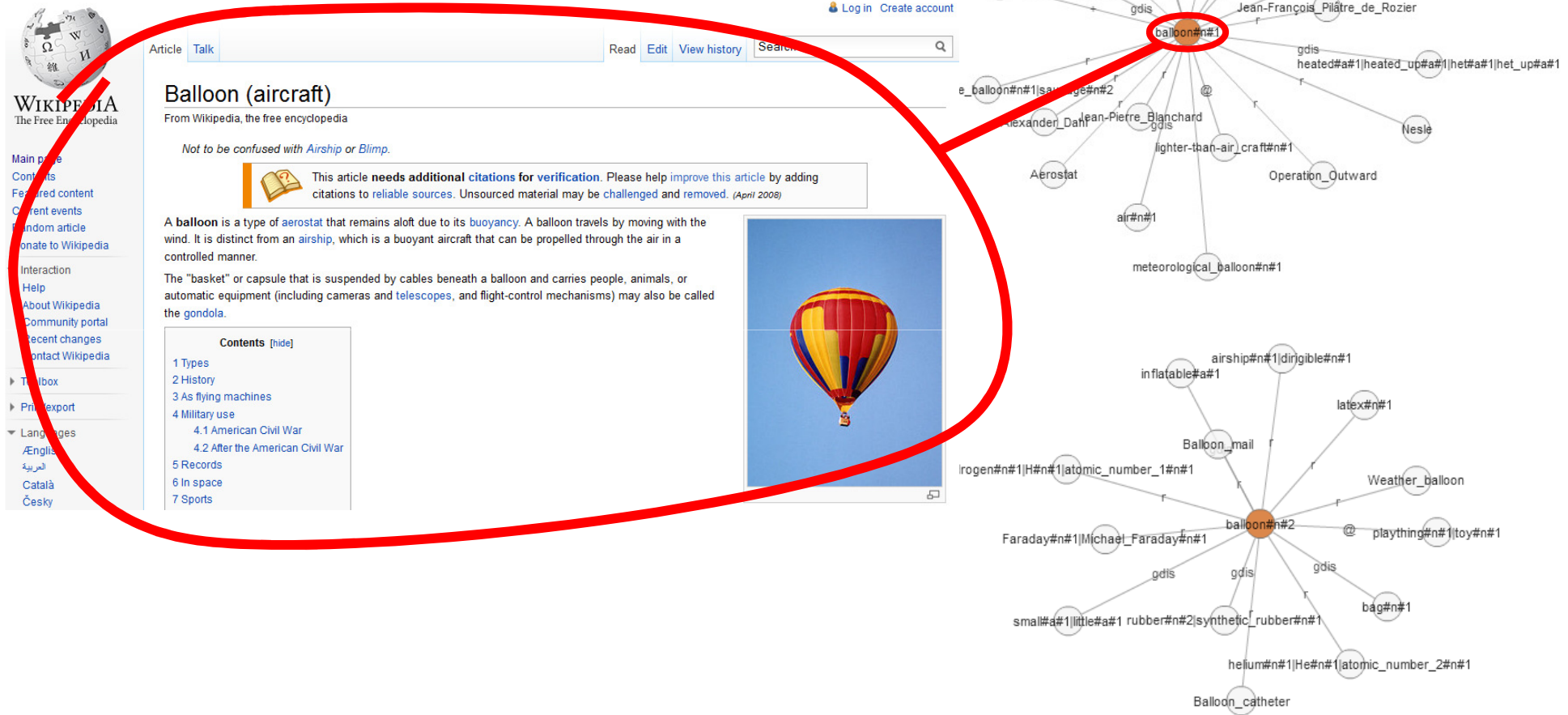
- By collecting the lexicalizations of concepts in different languages using:
 - a) Wikipedia interlanguage links
 - b) Statistical Machine Translation

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)

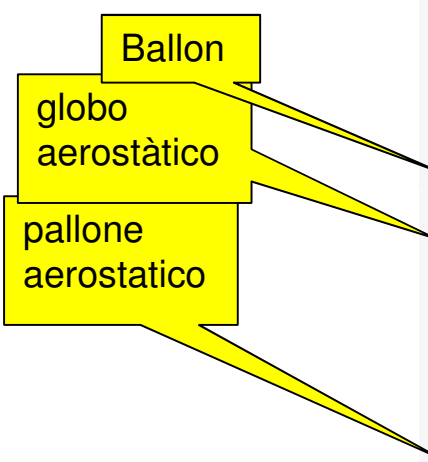
$$\begin{aligned}\mu(w) &= \underset{s \in Senses_{WN}(w)}{\operatorname{argmax}} p(s|w) = \underset{s}{\operatorname{argmax}} \frac{p(s, w)}{p(w)} \\ &= \underset{s}{\operatorname{argmax}} p(s, w)\end{aligned}$$

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

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

Building BabelNet: Translating Babel synsets

1. Exploiting Wikipedia interlanguage links



Ballon

globo aerostático

pallone aerostatico

About Wikipedia
Community portal
Recent changes
Contact Wikipedia

► Toolbox

► Print/export

▼ Languages

- /English
- العربية
- Català
- Česky
- Cymraeg
- Dansk
- Deutsch
- Eesti
- Ελληνικά
- Español
- Esperanto
- فارسی
- Français
- Frysk
- 한국어
- Hrvatski
- Bahasa Indonesia
- Íslenska
- Italiano
- עברית
- Қазақша
- Lietuvių
- മലയാളം
- 日本語
- Norsk (bokmål)
- Polski
- Português
- Română
- Русский

automatic equipment (including cameras and [telescopes](#), and night-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:

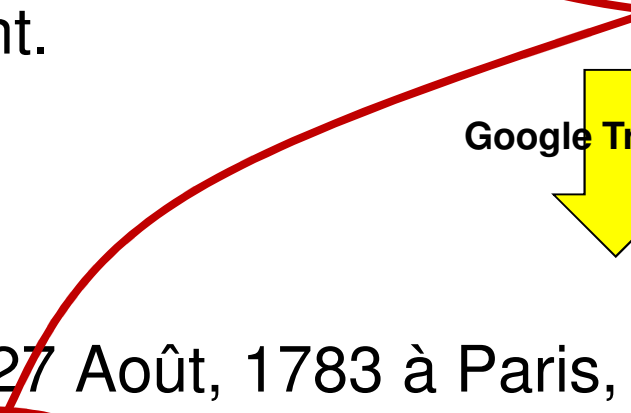
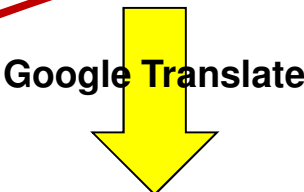
- **hot air balloons** obtain their buoyancy by heating the air inside the balloon. They are the most common type of balloon aircraft. "Hot air balloon" is sometimes used incorrectly to denote any balloon that carries people.
- **gas balloons** are inflated with a gas of lower **molecular weight** than 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.^[1]
- **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



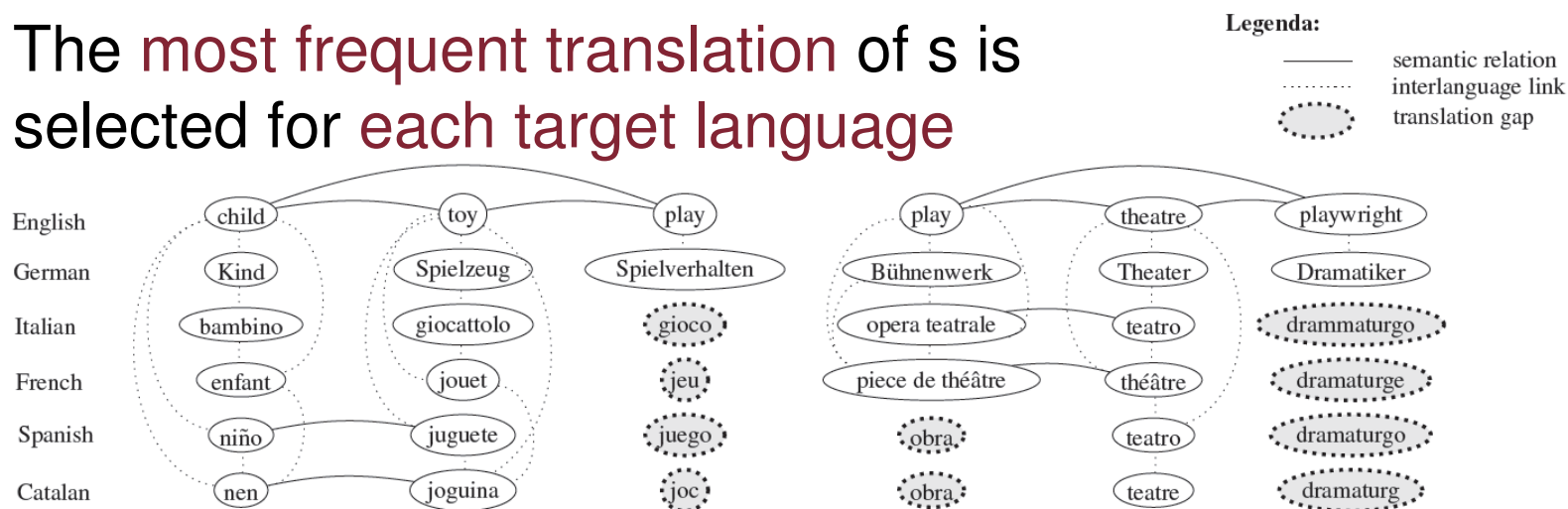
Building BabelNet: Translating Babel synsets

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.



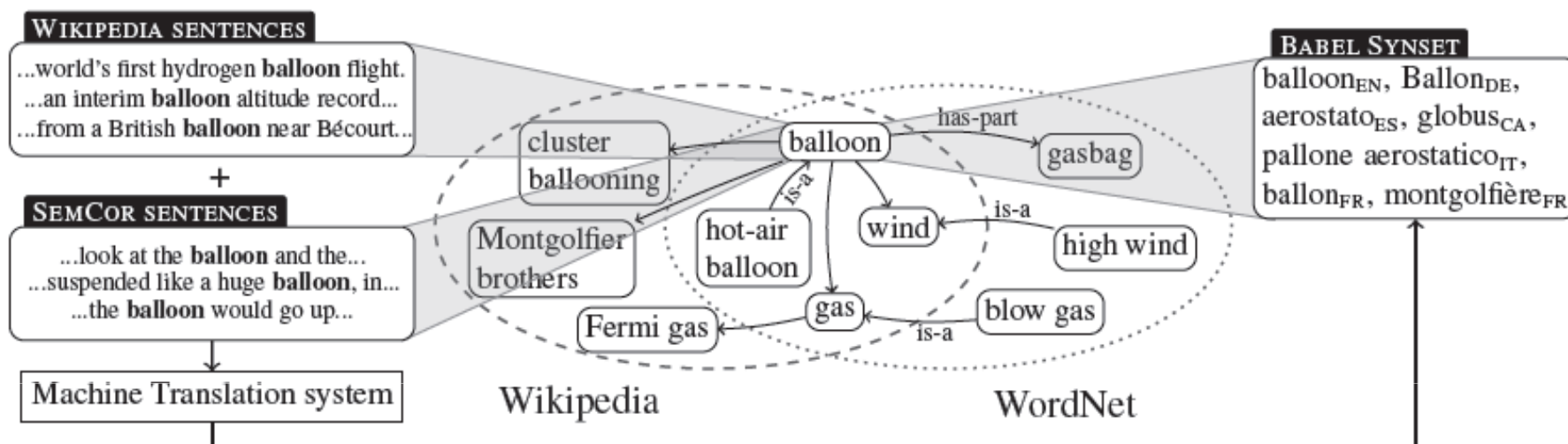
Google Translate
- Le 27 Août, 1783 à Paris, Franklin vu le premier vol en **ballon** d'hydrogène.

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 {	pages	123,101	524,897	506,892	404,153	349,375	4,863,970
	redirects	105,147	617,379	456,977	217,963	404,009	5,189,524
	translations	3,445,470	2,844,751	2,841,916	3,046,325	3,083,427	15,261,889
WordNet {	monosemous	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 ₁	A
BoW	taxonomic	89.7	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 $\left\{ \begin{array}{l} @ 2 \\ @ 3 \\ @ 4 \end{array} \right.$	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 $\left\{ \begin{array}{l} @ 2 \\ @ 3 \\ @ 4 \end{array} \right.$	80.5	60.6	69.1	77.0
		<u>77.5</u>	<u>65.2</u>	<u>70.9</u>	<u>78.2</u>
		<u>72.4</u>	<u>67.1</u>	<u>69.6</u>	<u>78.0</u>
	taxonomic + gloss relations				
	max depth $\left\{ \begin{array}{l} @ 2 \\ @ 3 \\ @ 4 \end{array} \right.$	<u>81.2</u>	<u>74.6</u>	<u>77.7</u>	<u>82.7</u>
		72.8	77.4	75.1	80.1
		64.3	76.2	69.8	75.0
	MFS baseline	25.4	49.2	33.5	25.4
	Random baseline	24.2	46.9	31.9	24.2

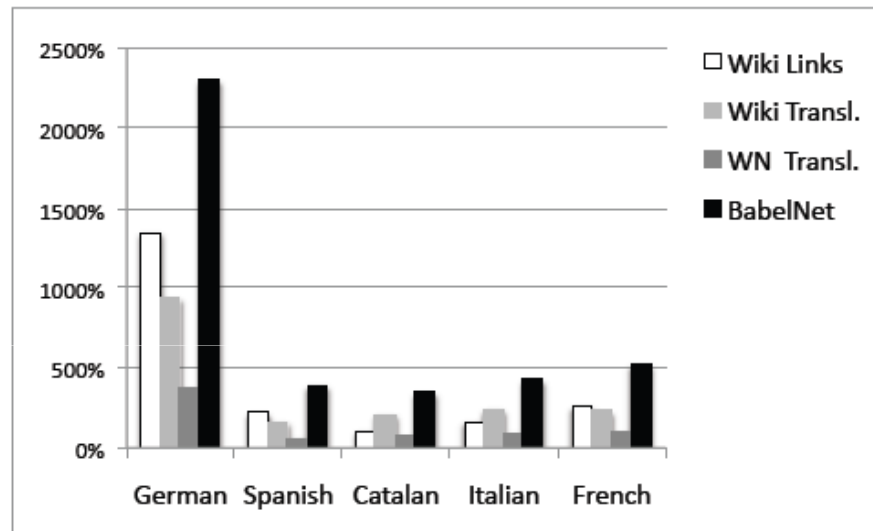
Evaluation of BabelNet against gold standard resources

Coverage

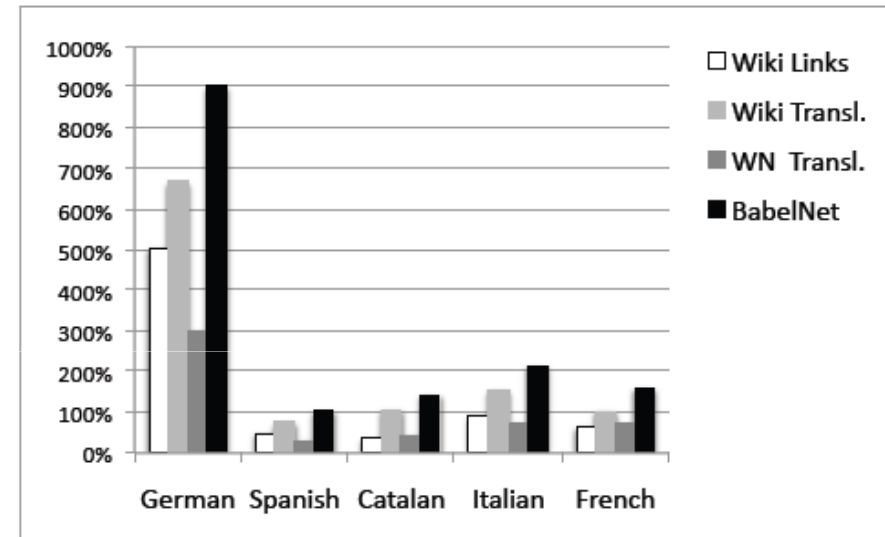
	Resource	Method	SENSES	SYNSETS
German	WIKI	Links	39.6	50.7
		Transl.	42.6	58.2
	WN	Transl.	21.0	28.6
	BABELNET	All	57.6	73.4
Spanish	WIKI	Links	34.4	40.7
		Transl.	47.9	56.1
	WN	Transl.	25.2	30.0
	BABELNET	All	66.4	76.6
Catalan	WIKI	Links	20.3	25.2
		Transl.	46.9	54.1
	WN	Transl.	25.0	29.6
	BABELNET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

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 ₁	All words P/R/F ₁
WordNet	Degree	80.1	79.7
	PLength	80.2	79.8
	SProbability	79.8	79.3
	PageRank	79.9	79.4
BabelNet	Degree	84.7	82.3
	PLength	85.4	82.7
	SProbability	84.6	82.1
	PageRank	82.3	80.1
	SUSSX-FR	81.1	77.0
	TreeMatch	N/A	73.6
	NUS-PT	82.3	82.5
	SSI	84.1	83.2
	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 {	837,627	82,115	845,210
	–	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/bnexplorer>

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 displays the BabelNetXplorer web application. On the right side, the 'Input Words' field contains the word 'bank'. A yellow callout box labeled 'Input word' points to this field. Below the input field, the 'Knowledge Base' is set to 'BabelNet'. The search results are displayed in a list, starting with the word 'bank' followed by its BabelNet ID 'bn:00008363n'. The results are categorized by part of speech, with 'NOUN' being the first category. The list includes various senses of the word 'bank', such as 'depository financial institution', 'savings bank', 'bank building', and 'Bank (tobacco)'. The interface also features a 'Find' field and 'Search' and 'Filters' buttons at the bottom left. The BabelNet logo is visible in the top right corner, and the Linguistic Computing Laboratory logo is in the bottom right corner.

BabelNetXplorer

Input Words: Knowledge Base: BabelNet

Search

NOUN

bank#n#1 bn:00008363n
CA:vora, CA:riba, IT:riva, ES:orilla
FR:rive, DE:bank
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
bank#n#9|bank building#n#1 bn:00008371n
bank#n#10 bn:00008372n
Bank (tobacco) bn:00919659n
Bank (topography) bn:00932118n
CA:banc, IT:banca, ES:banco, FR:banque

Find: Search Filters

Linguistic Computing Laboratory

BabelNetXplorer: semantic network exploration

- Click a **Babel** sense of the input word:

Selected sense

The screenshot displays the BabelNetXplorer interface. On the left, a semantic network for the word 'bank' is shown, with a central node 'bank' connected to various related terms like 'financial_services', 'investment_banking', 'branch_(banking)', 'thrift_institution', 'central_bank', 'member_bank', 'interest', 'state_bank', 'finance', 'currency', 'interest_rate', 'rate_of_interest', 'bankruptcy', 'deposit', 'banking_concern', 'banking_company', 'banking_industry', 'banking_system', 'savings_and_loan', 'merchant_bank', 'insurance', 'double-entry_bookkeeping_system', 'financial_institution', 'financial_organization', 'financial_organisation', 'commercial_bank', 'full_service_bank', 'retail_banking', 'transactional_account', 'credit_card', 'charge_card', 'charge_plate', 'plastic', 'credit_union', 'bill', 'note', 'government_note', 'bank_bill', 'banker's_bill', 'bank_note', 'banknote', 'Federal_Reserve_note', 'greenback', 'check', 'bank_check', 'cheque', 'channel', 'canalized', 'agent_bank', 'building_society', 'acquirer', 'cash_machine', 'cash_dispenser', 'automated_teller_machine', 'automatic_teller_machine', 'ATM', 'bond', 'bond_certificate', 'accept', 'take', 'have', 'bond', 'bond_certificate', 'accept', 'take', 'have'. On the right, the 'BabelNetXplorer' interface is shown with the input word 'bank' and a list of senses. A yellow callout points to the selected sense 'bank#n#1'.

BabelNetXplorer

Input Words: bank

Knowledge Base: BabelNet

Search

NOUN

bank#n#1 bn:00008363n
CA.vora, CA.riba, IT.riva, ES.orilla
FR.rive, DE.bank

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

bank#n#9/bank building#n#1 bn:00008371n

bank#n#10 bn:00008372n

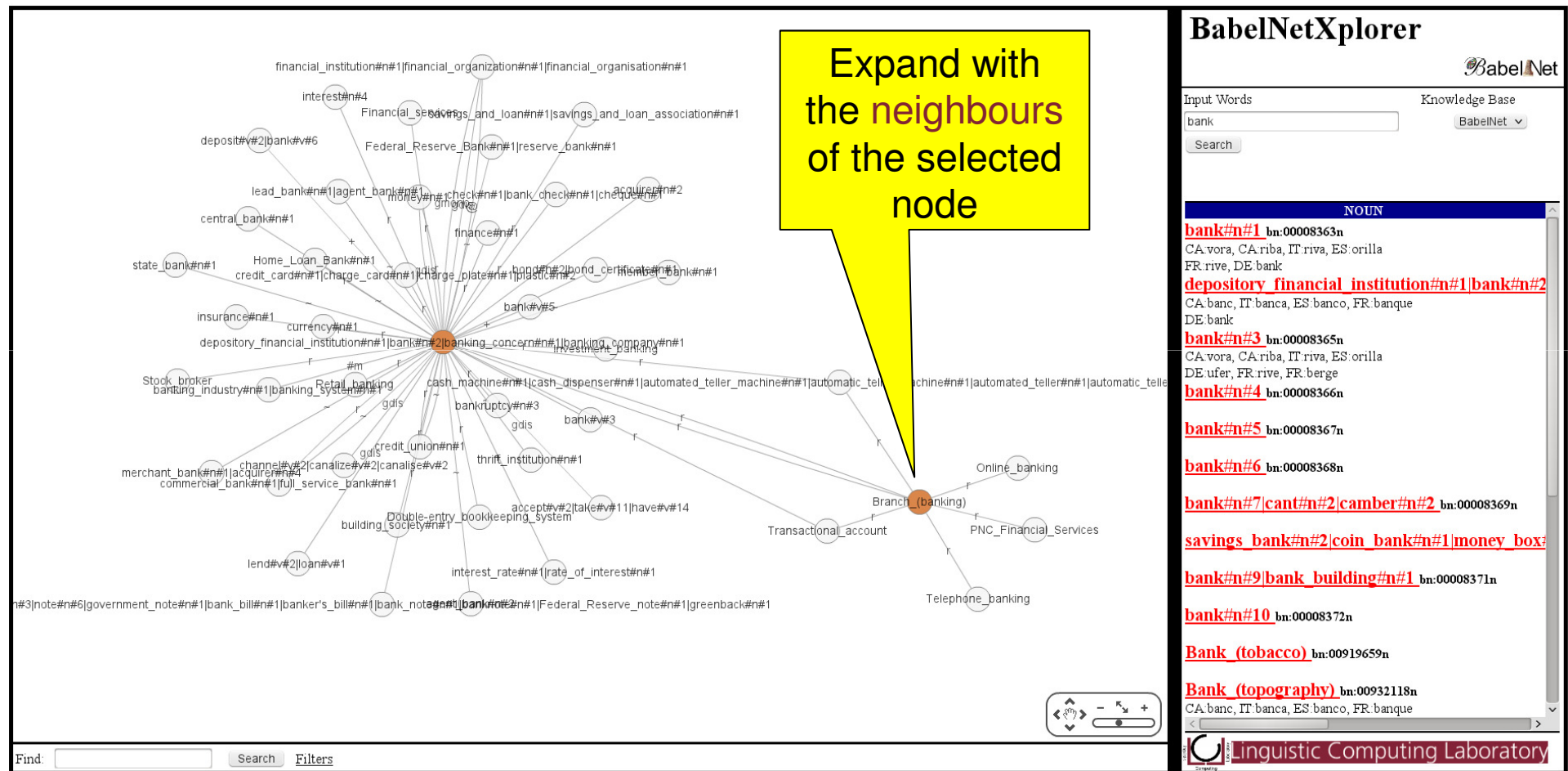
Bank (tobacco) bn:00919659n

Bank (topography) bn:00932118n
CA.banc, IT.banca, ES.banco, FR.banque

Find: Search Filters

Linguistic Computing Laboratory

- Expand the graph by clicking on a node:



BabelNetXplorer: semantic network exploration

- Expand the graph by clicking on a node:

The screenshot displays the BabelNetXplorer interface. On the left, a complex semantic network graph is shown with nodes representing various concepts related to banking and finance. A yellow callout box points to a node in the graph, stating: "Expand with the neighbours of the selected node". On the right, the BabelNetXplorer interface is shown, featuring a search bar with the input "bank" and a dropdown menu for the knowledge base. Below the search bar, a list of related terms is displayed, including "bank#n#1", "depository financial institution#n#1", "bank#n#2", "bank#n#3", "bank#n#4", "bank#n#5", "bank#n#6", "bank#n#7", "cambor#n#2", "camber#n#2", "savings bank#n#2", "coin bank#n#1", "money box", "bank#n#9", "bank building#n#1", "bank#n#10", "Bank (tobacco)", and "Bank (topography)". The interface also includes a "Find" search bar at the bottom left and a "Linguistic Computing Laboratory" logo at the bottom right.

BabelNetXplorer

Input Words: bank Knowledge Base: BabelNet

Search

NOUN

bank#n#1 bn:00008363n
CA.vora, CA.riba, IT.riva, ES.orilla
FR.rive, DE.bank

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

bank#n#9|bank building#n#1 bn:00008371n

bank#n#10 bn:00008372n

Bank (tobacco) bn:00919659n

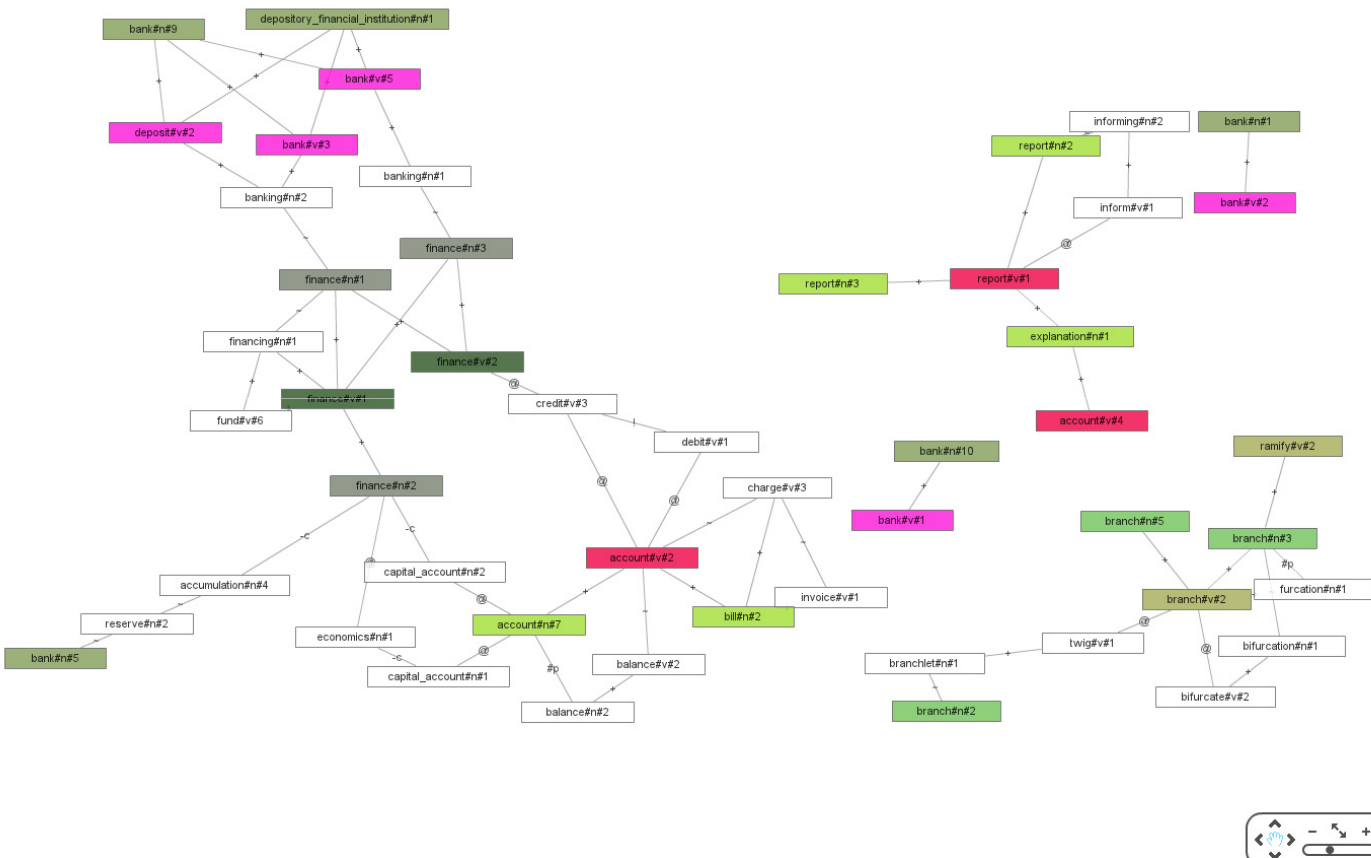
Bank (topography) bn:00932118n
CA.banc, IT.banca, ES.banco, FR.banque

Find: Search Filters

Linguistic Computing Laboratory


BabelNetXplorer: search for connecting paths

- Search the graph for connecting paths:



Find: Search Filters

BabelNetXplorer



Input Words: Knowledge Base:

Search

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!

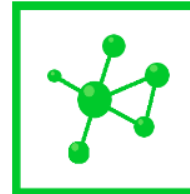
Create a disambiguation
graph for the target words
And disambiguate in 1 line!

Coming soon to your screens: BabelNet 1.1!

BabelNet^{1.1}
A very large multilingual ontology

means: 40 languages +
more accurate mappings
and translations!

search **disambiguate**



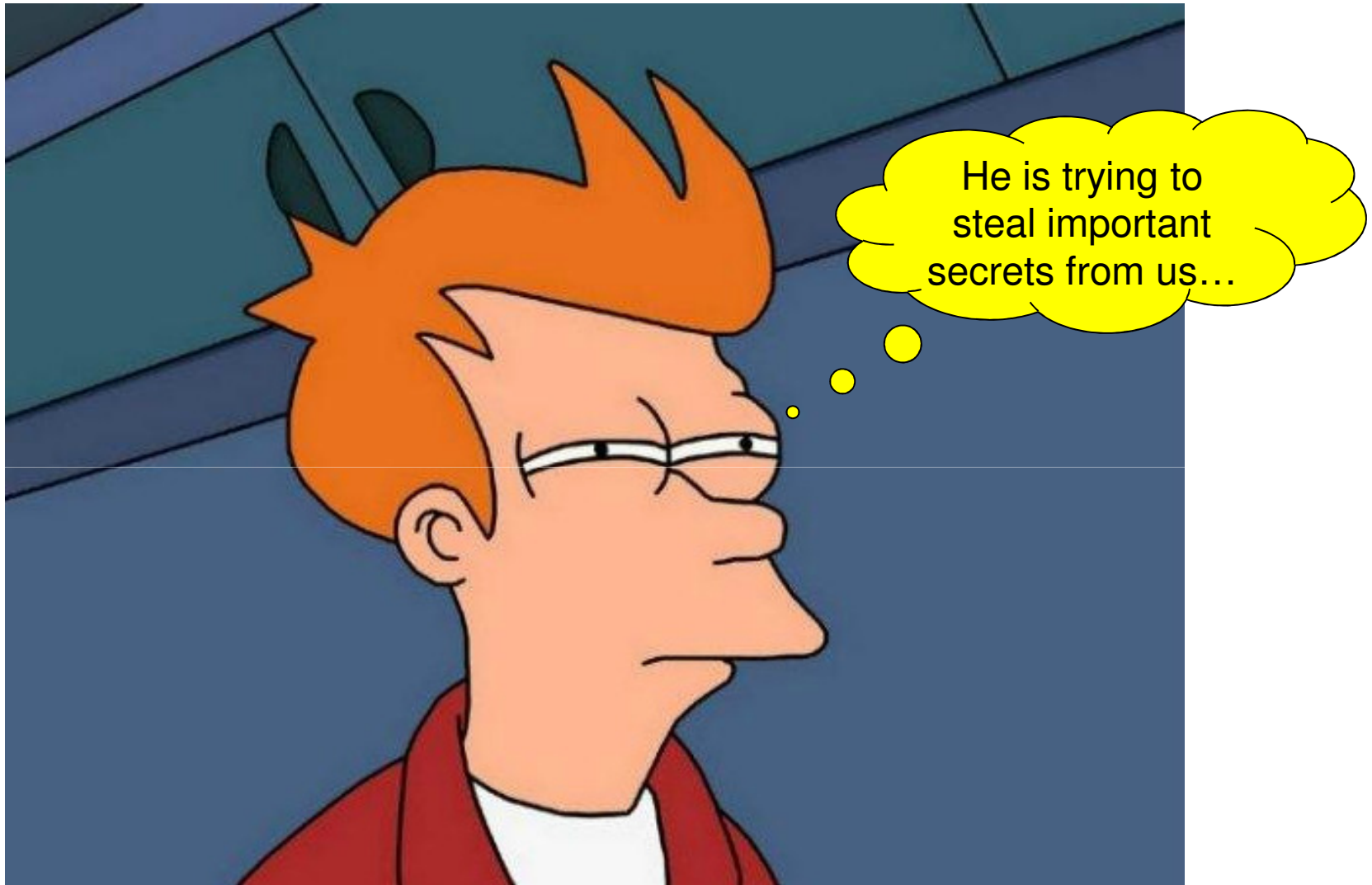
PUBLICATIONS **download**



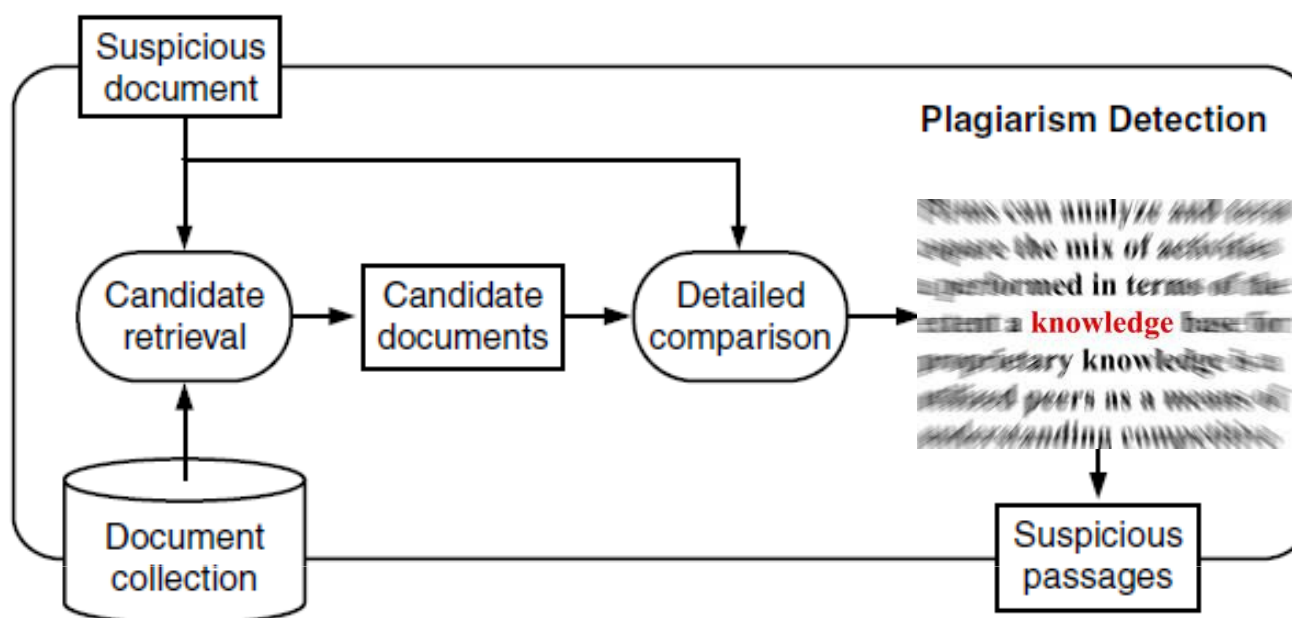
BabelNet is an output of the ERC Starting Grant MultiJEDI No. 259234.



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

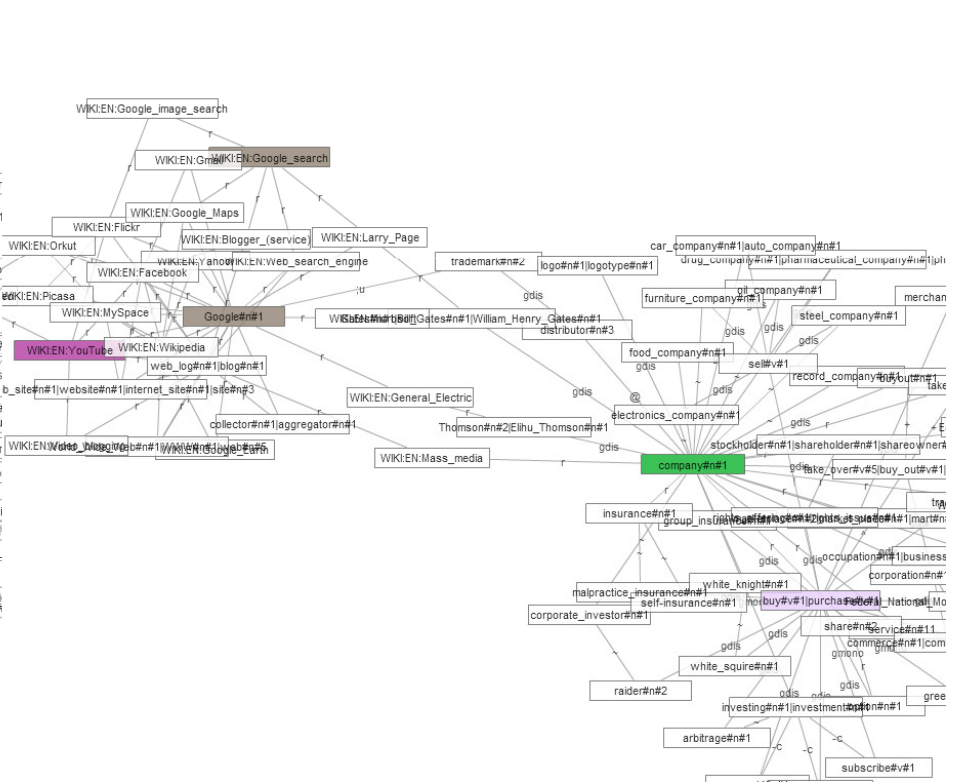


Plagiarism detection: the state of the art

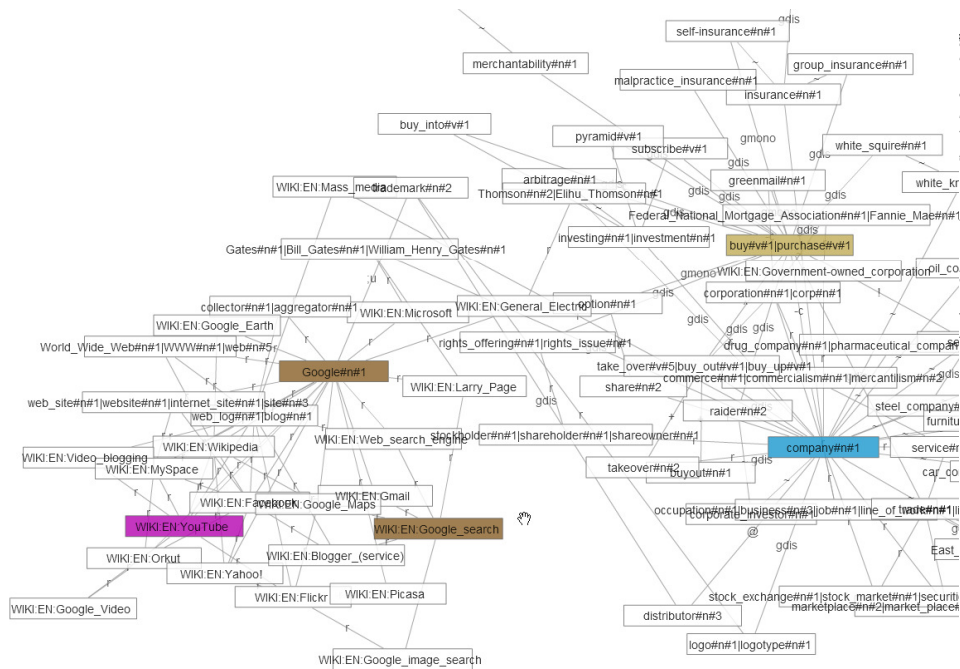


- Stemming, stopword removal, chunking into passages, keyphrase extraction, n-grams, query formulation, search control, etc.

Google bought YouTube



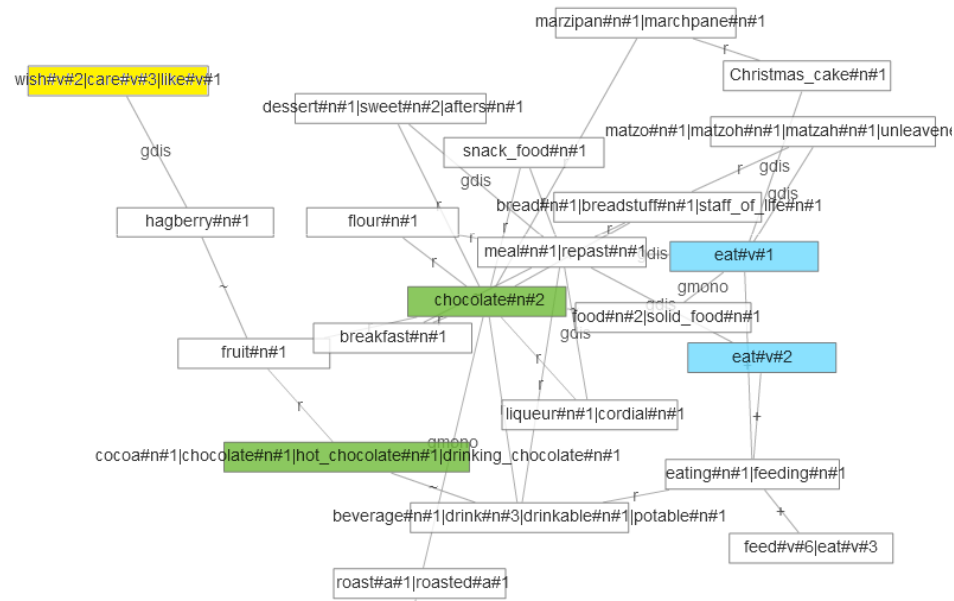
Google bought YouTube



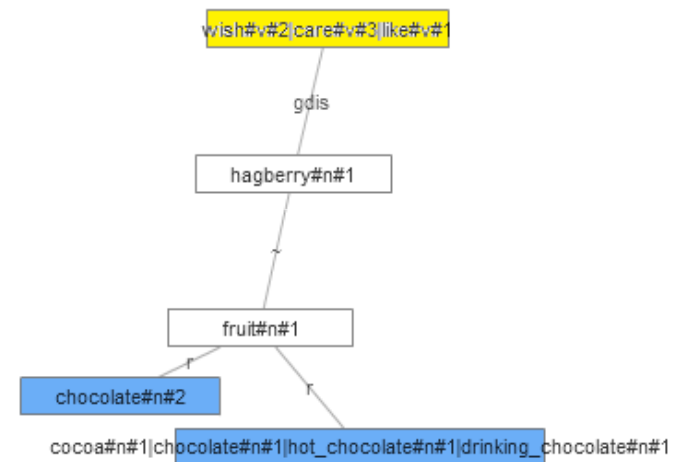
So, what can we do?

- **Deletion:**

I like eating chocolate



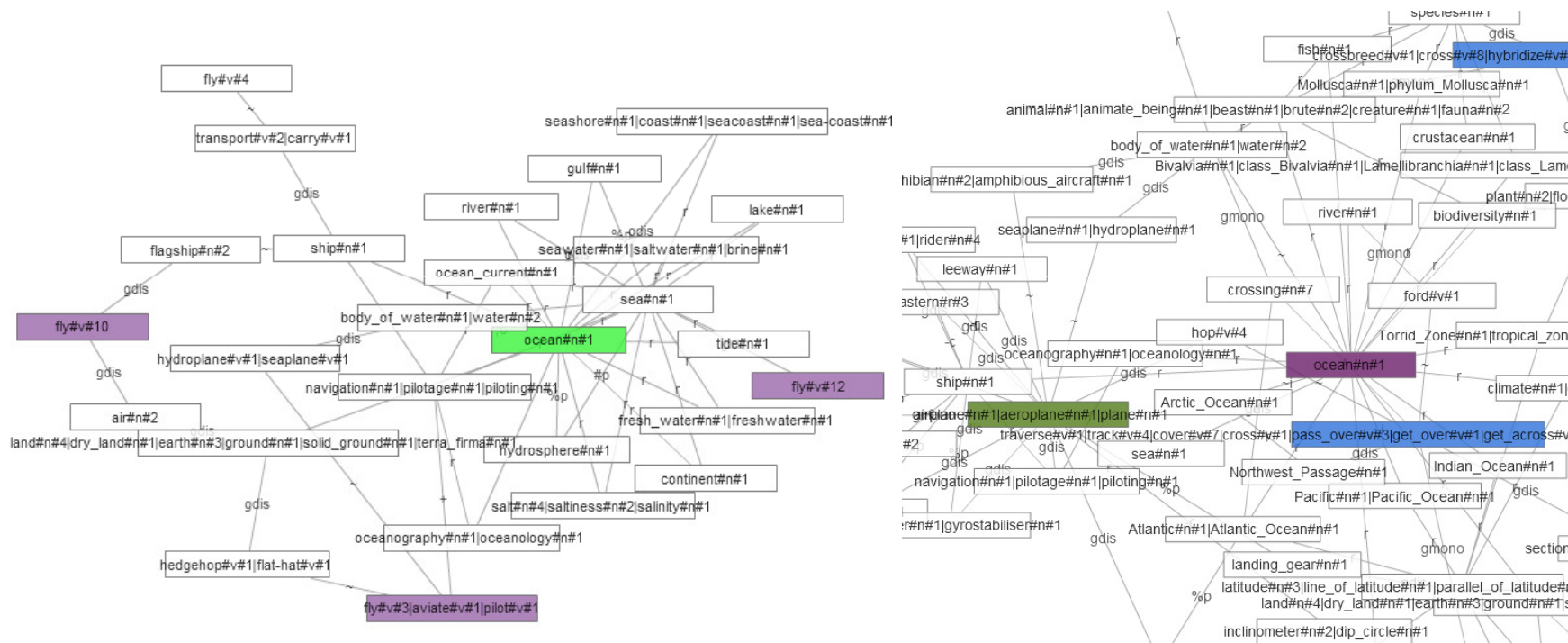
I like chocolate



- **Semantics based changes:**

Bill flew across the ocean

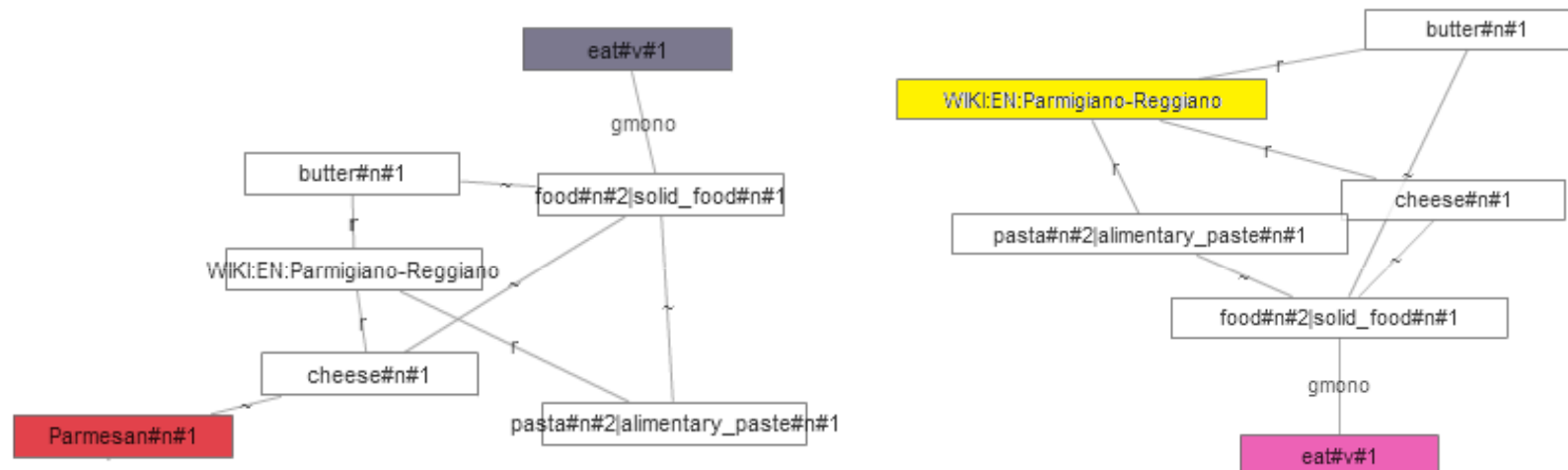
Bill crossed the ocean by plane



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.  
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

