

# Multilingual Word Sense Disambiguation and Entity Linking

Roberto Navigli

DIPARTIMENTO  
DI INFORMATICA



SAPIENZA  
UNIVERSITÀ DI ROMA

Linguistic Computing Laboratory

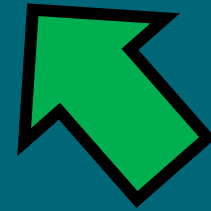
<http://lcl.uniroma1.it>

ERC Starting Grant n. 259234

Lisbon, 22<sup>nd</sup> July 2015

# Multilingual Word Sense Disambiguation and Entity Linking [with BabelNet & Babelfy]

Roberto Navigli



DIPARTIMENTO  
DI INFORMATICA



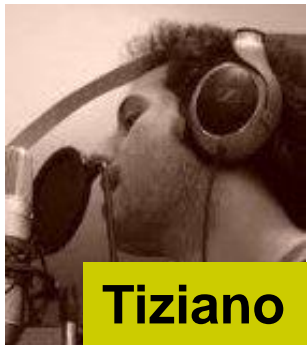
SAPIENZA  
UNIVERSITÀ DI ROMA

Linguistic Computing Laboratory

<http://lcl.uniroma1.it>

ERC Starting Grant n. 259234

Lisbon, 22<sup>nd</sup> July 2015

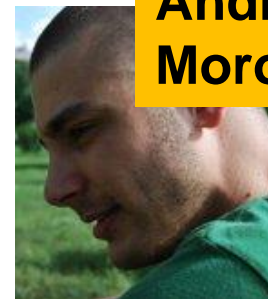


**Tiziano  
Flati**



**Daniele  
Vannella**

**Alessandro  
Raganato**



**Andrea  
Moro**



**Simone Ponzetto**



**Francesco  
Cecconi**



**Taher Pilehvar**



**Federico  
Scozzafava**



**Ignacio Iacobacci**



**José  
Camacho  
Collados**

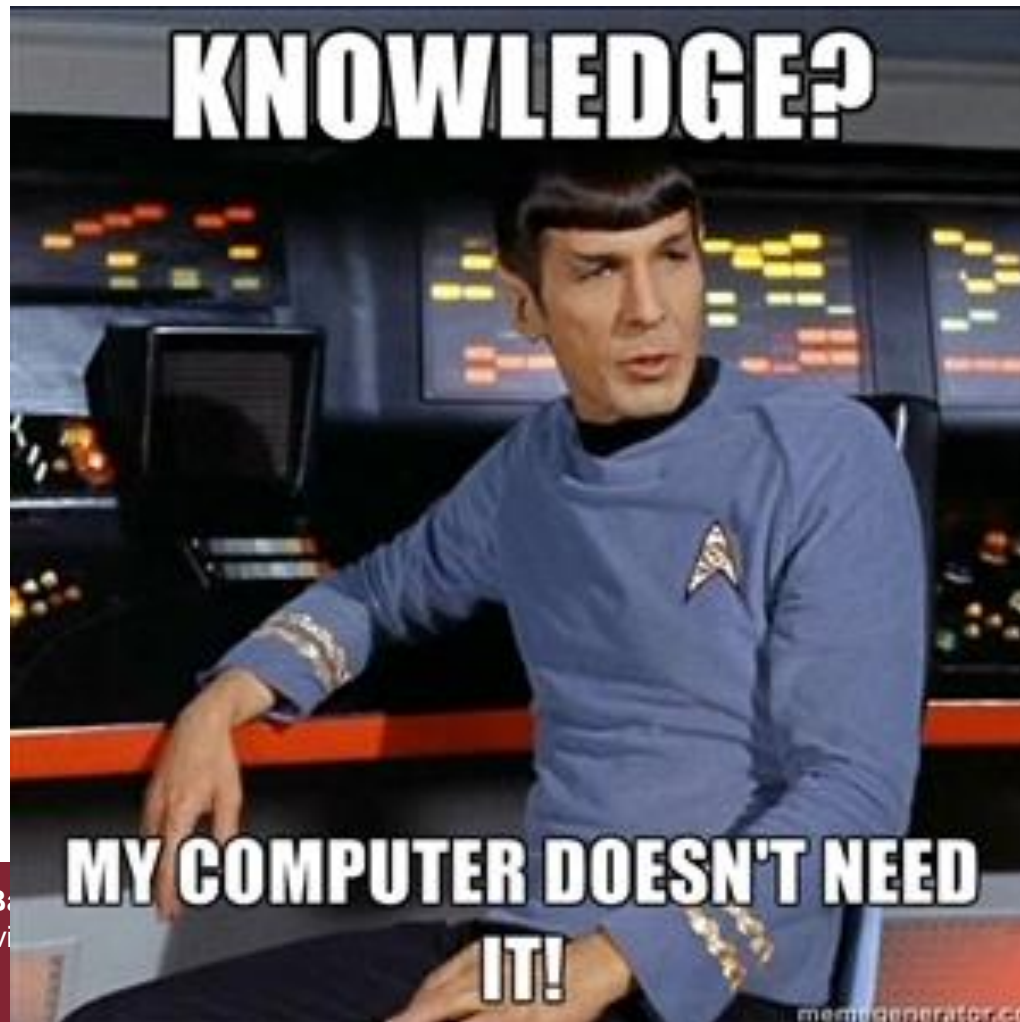
27/07/2015



**knowledge**

# It's all about knowledge!

- But can we expect computers to **know**?
- Can't computers just use, e.g., **statistical techniques**?



# State-of-the-art Machine Translation

- **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).



# State-of-the-art Machine Translation

- **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).
- **ES:** Estas son las películas en las que el género de la música, por ejemplo, **roca**, es un elemento importante, pero no necesariamente el centro de la trama. [...]



# State-of-the-art Machine Translation

- **EN:** We can look at how this vast slug of molten underground **rock** was injected.

Danger here!



# State-of-the-art Machine Translation

- **EN:** We can look at how this vast slug of molten underground **rock** was injected.
- **FR:** Nous pouvons voir comment ce vaste bouchon de **rock** underground fondu a été injecté.
- **IT:** Possiamo guardare a come è stato iniettato questo vasto slug del **rock** underground fusa.

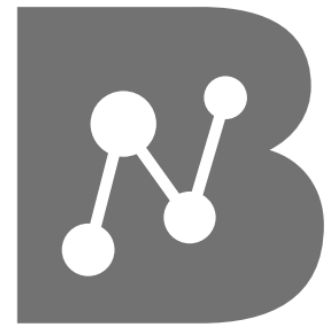


# What are we talking about?



A stylized, geometric representation of the word 'JSD' (Joint word sense Disambiguation) using black lines and dots.  
Multilingual Joint word sense Disambiguation ■

A **5-year ERC Starting Grant** (2011-2016)  
on Multilingual Word Sense Disambiguation



BabelNet

# INTEGRATING KNOWLEDGE

[Navigli & Ponzetto, ACL 2010;  
Pilehvar & Navigli, ACL 2014]

# The resource diaspora

WordNet is a multilingual free encyclopedia

Wiktionary [ˈwɪkʃənəri] n., a wiki-based Open Content dictionary

Word to search Wikia [ˈwiːkɪə]

Display Options

Key: "S:" =

Display options

Noun

• S:



Meta:Main Page  
Visual Dictionary  
Random expression  
Recent changes  
OmegaWiki blog

Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia  
Wikipedia store

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact page

Tools  
What links here  
Related changes

English [edit]

Etymology [edit]

From *compute* + *-er*.

Pronunciation [edit]

- (UK) IPA<sup>(key)</sup>: /kəmˈpjʊ:tə/

- Audio (UK) 0:00 MENU

- (US) IPA<sup>(key)</sup>: /kəmˈpjʊtə/

- Audio (US) 0:00 MENU

- Hyphenation: com·put·er

- Rhymes: -u:tə(r)

Expression Discussion

Read Edit View history



Search

As an anonymous user, you can only add new data. If you would like to also modify existing data, please create an account and indicate your languages on your user page.

machine

Other languages: Dutch French

Language: English

Substantive

▼ **machine** : A device able to perform a particular, more or less complex, job. [Edit]

▼ Lexical annotations

Property	Value
----------	-------

word class	substantive
------------	-------------

▼ Definition

Language	Text
----------	------

Castilian	Dispositivo capaz de realizar un trabajo particular, más o menos complejo.
-----------	--

English	A device able to perform a particular, more or less complex, job.
---------	---

more complex machines exist. Examples include [vehicles](#), [electronic systems](#), [molecular machines](#), [computers](#), [television](#), and [radio](#).

Contents [hide]

1 Etymology

2 History

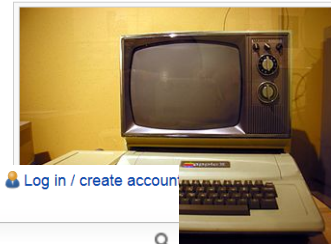
3 Types



Wikimedia Commons has related media at:  
[computer](#)



Wikipedia has an article on:  
[Computer](#)



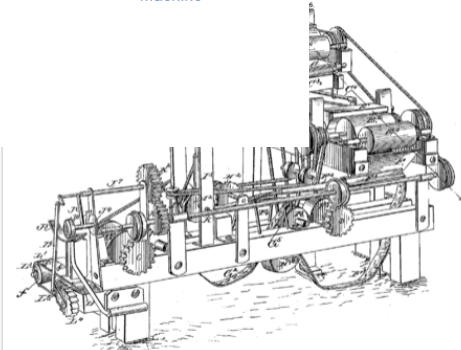
English Log in / create account

ater (circa early

retrieve large  
Internet, or

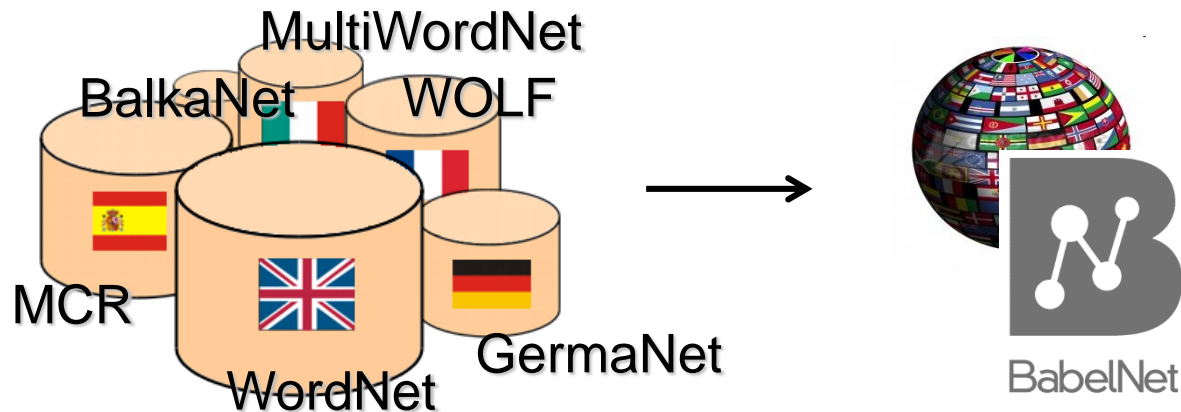


Machine



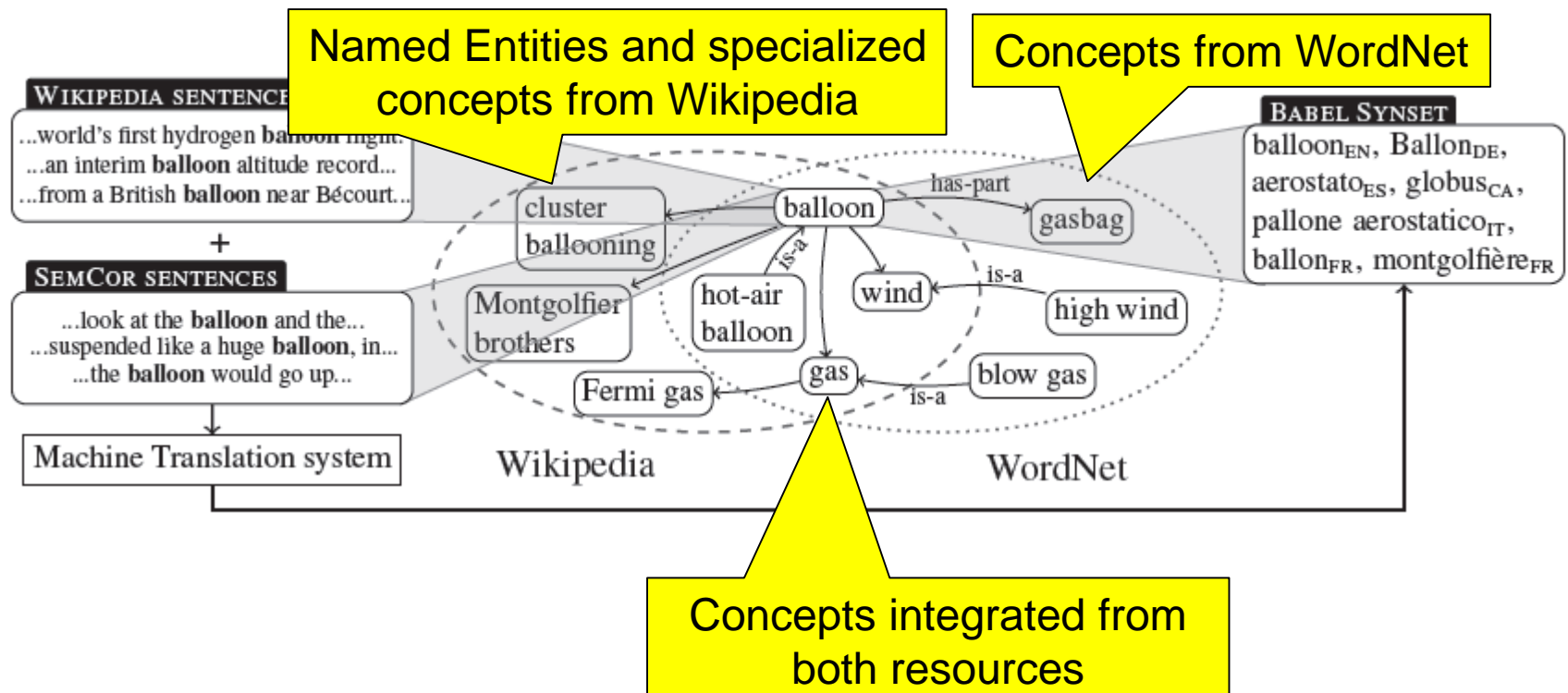
# Multilingual Joint Word Sense Disambiguation (MultiJEDI)

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



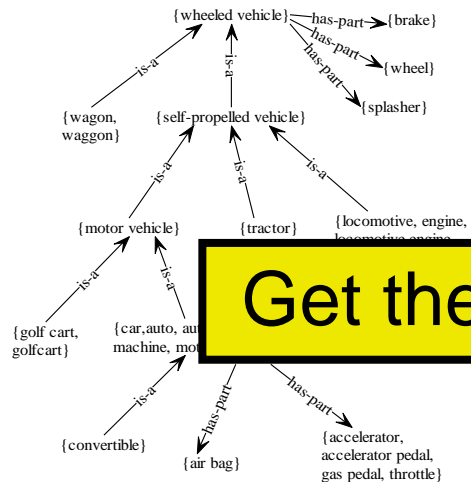
# It all started with merging WordNet and Wikipedia [Navigli and Ponzetto, ACL 2010; AIJ 2012]

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

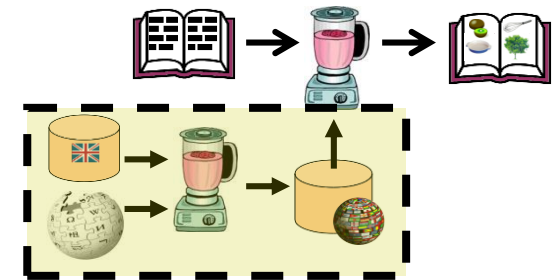


# Creating a Multilingual Semantic Network

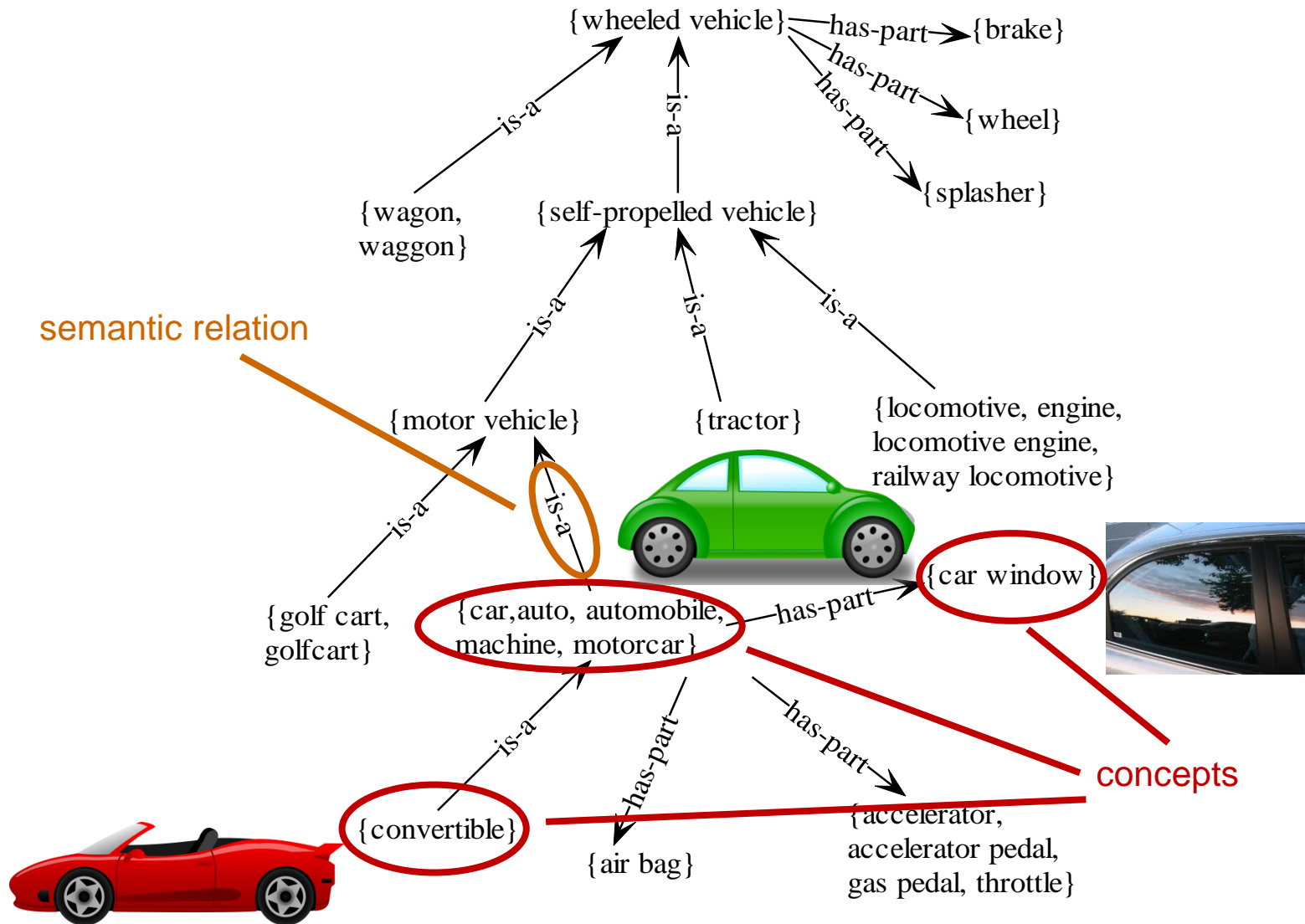
- Start from two large **complementary** resources:
  - **WordNet**: full-fledged taxonomy
  - **Wikipedia**: multilingual and continuously updated



Get the best from both worlds



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



# Wikipedia [The Web Community, 2001-today]

## Automobile

From Wikipedia, the free encyclopedia  
(Redirected from [Car](#))

For the magazine, see [Automobile Magazine](#).

"[Car](#)" redirects here. For other uses, see [Car \(disambiguation\)](#).

An **automobile**, **autocar**, **motor car** or **car** is a wheeled [motor vehicle](#) used for [transporting passengers](#), which also carries its own engine or motor. Most definitions of the term specify that automobiles are designed to run primarily on roads, to have seating for one to eight people, to typically have four wheels, and to be constructed principally for the transport of people rather than goods.<sup>[3]</sup>

The term *motorcar* has also been used in the context of electrified rail systems to denote a car which functions as a small locomotive but also provides space for passengers and baggage. These locomotive cars were often used on suburban routes by both interurban and intercity railroad systems.<sup>[4]</sup>

It was estimated in 2010 that the number of automobiles had risen to over 1 billion vehicles, with 500 million reached in 1986.<sup>[5]</sup> The numbers are increasing rapidly, especially in [China](#) and [India](#).<sup>[6]</sup>



## Motor vehicle

From Wikipedia, the free encyclopedia

A **motor vehicle** or **road vehicle** is a self-propelled wheeled [vehicle](#) that does not operate on rails, such as [trains](#) or [trolleys](#). The [vehicle propulsion](#) is provided by an [engine](#) or motor, usually by an [internal combustion engine](#), or an [electric motor](#), or some combination of the two, such as [hybrid electric vehicles](#) and [plug-in hybrids](#). For legal purposes motor vehicles are often identified within a number of vehicle classes including [automobiles](#) or cars, [buses](#), [motorcycles](#), [motorized bicycles](#), [off highway vehicles](#), [light trucks](#) or light duty trucks, and [trucks](#) or lorries. These classifications vary according to the legal codes of each country. ISO 3833:1977 is the standard for road vehicles types, terms and definitions.<sup>[1]</sup>

As of 2010 there were more than one billion motor vehicles in use in the world excluding [off-road vehicles](#) and [heavy construction equipment](#).<sup>[2][3][4]</sup> Global vehicle ownership [per capita](#) in 2010 was 148 vehicles in operation per 1000 people.<sup>[4]</sup> The United States has the largest fleet of motor vehicles in the world, with 239.8 million by 2010. Vehicle



(unspecified) semantic relation

concepts

## Passenger

From Wikipedia, the free encyclopedia

*This article is about passengers in commercial transportation; for other uses see [Passenger \(disambiguation\)](#)*

A **passenger** is a person who travels in a [vehicle](#) but bears little or no responsibility for the tasks required for that vehicle to arrive at its destination or otherwise operate the vehicle.

Passengers are people who ride on [buses](#), [passenger trains](#), [airliners](#), [ships](#), [ferryboats](#), and other methods of transportation.

Crew members (if any), as well as the driver or pilot of the vehicle, are considered to be passengers. For example, a [flight attendant](#) would not be considered a "passenger" while on duty, but an [idling](#) in a [company car](#) being driven by another person would be a [passenger](#), even if the car was being driven in company

Look up [passenger](#) in Wiktionary, the free dictionary.



## Travel

From Wikipedia, the free encyclopedia  
(Redirected from [Traveling](#))

For other uses, see [Travel \(disambiguation\)](#).

**Travel** is the movement of [people](#) or objects (such as [airplanes](#), [boats](#), [trains](#) and other conveyances) between relatively distant geographical [locations](#).<sup>[1][2]</sup>

**Contents** [hide]

- 1 Etymology
- 2 Purpose and motivation
- 3 Travel safety
- 4 See also
- 5 References
- 6 External links

## Etymology

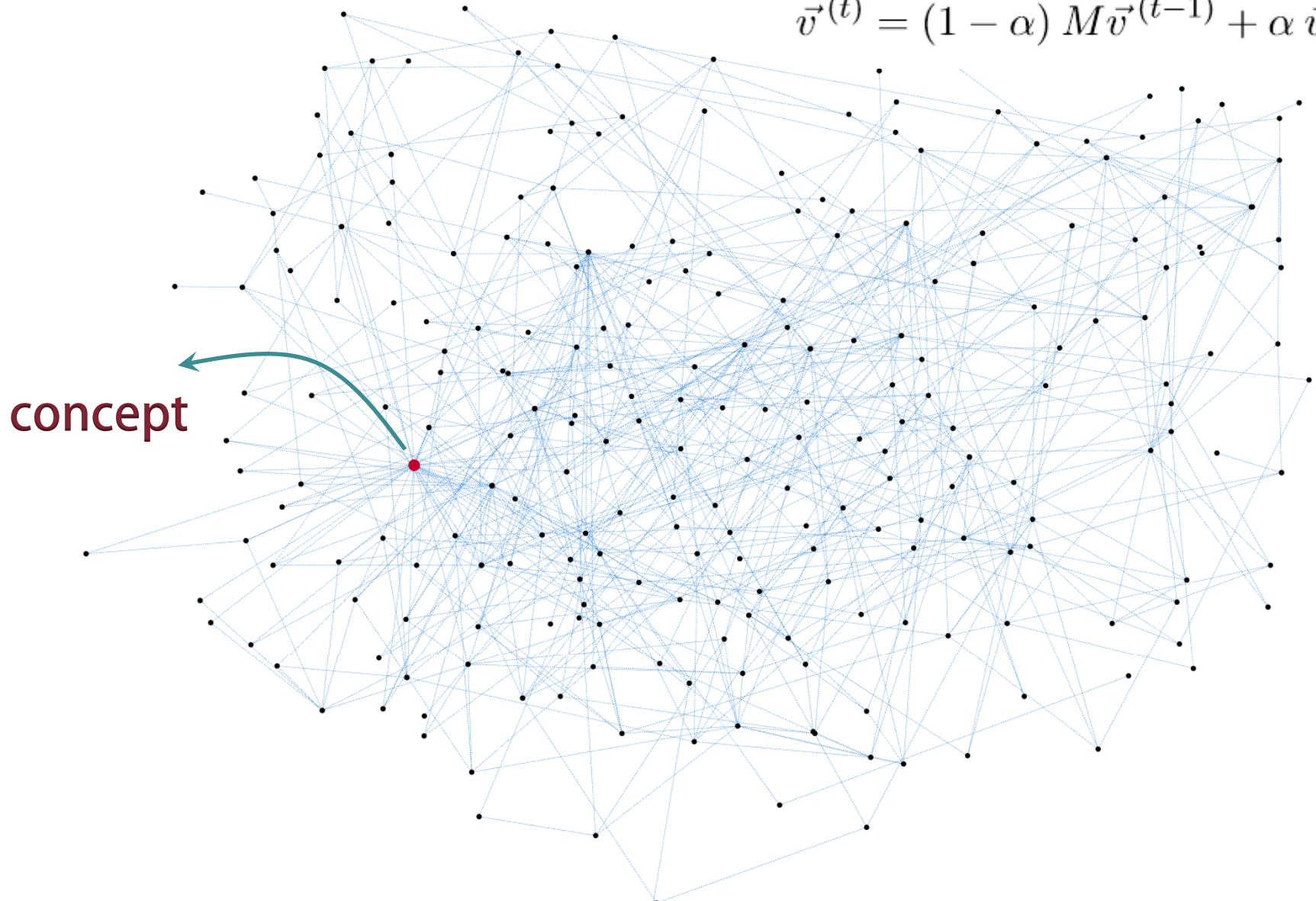
The term "travel" originates from the Old French word *travail*.<sup>[3]</sup> The term also covers all the activities performed during a travel (movement).<sup>[4]</sup>



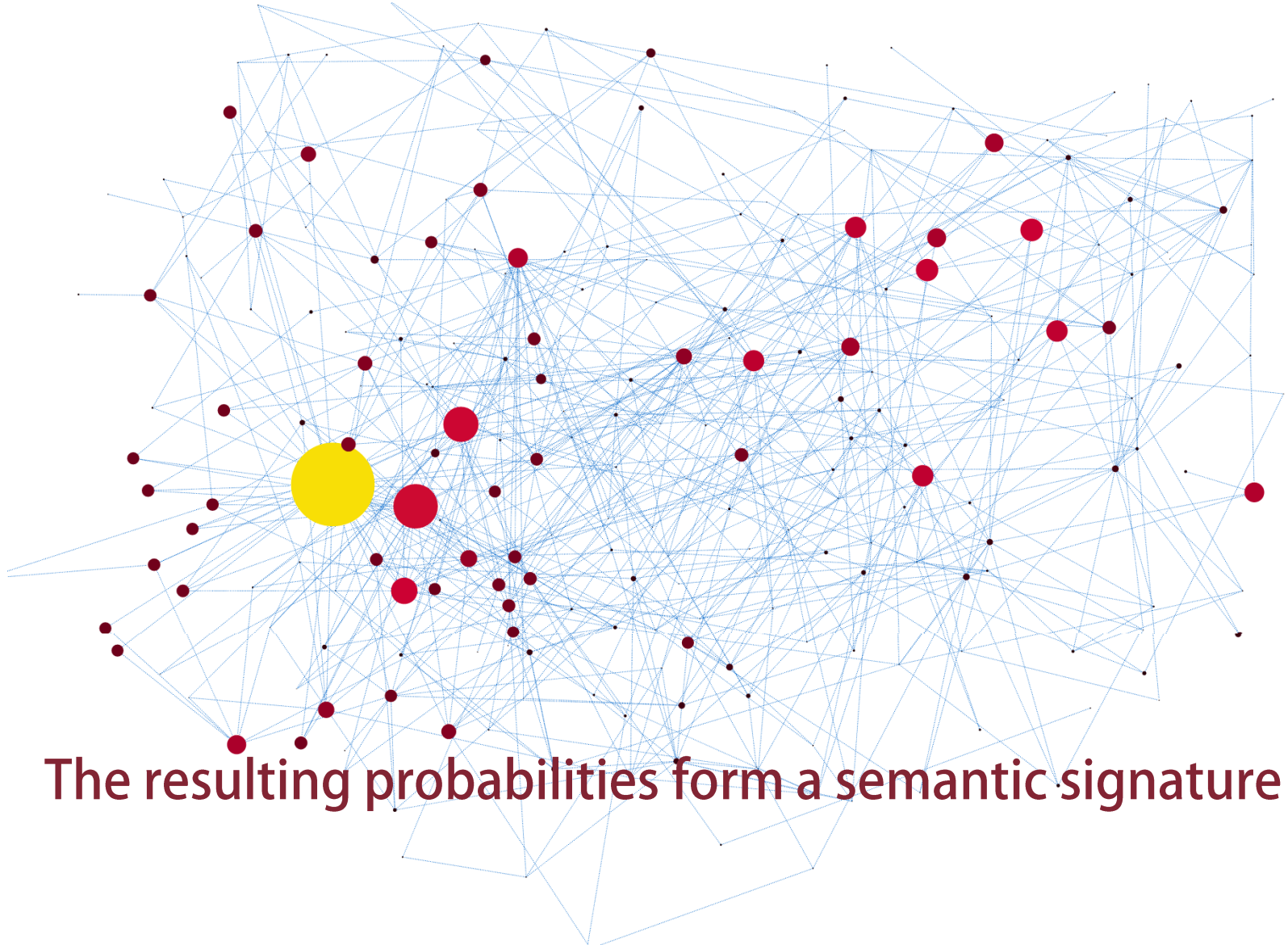
Roberto Navigli

# Structural Similarity with Personalized PageRank [Pilehvar and Navigli, ACL 2014]

$$\vec{v}^{(t)} = (1 - \alpha) M \vec{v}^{(t-1)} + \alpha \vec{v}^{(0)}$$



# Structural Similarity with Personalized PageRank [Pilehvar and Navigli, ACL 2014]

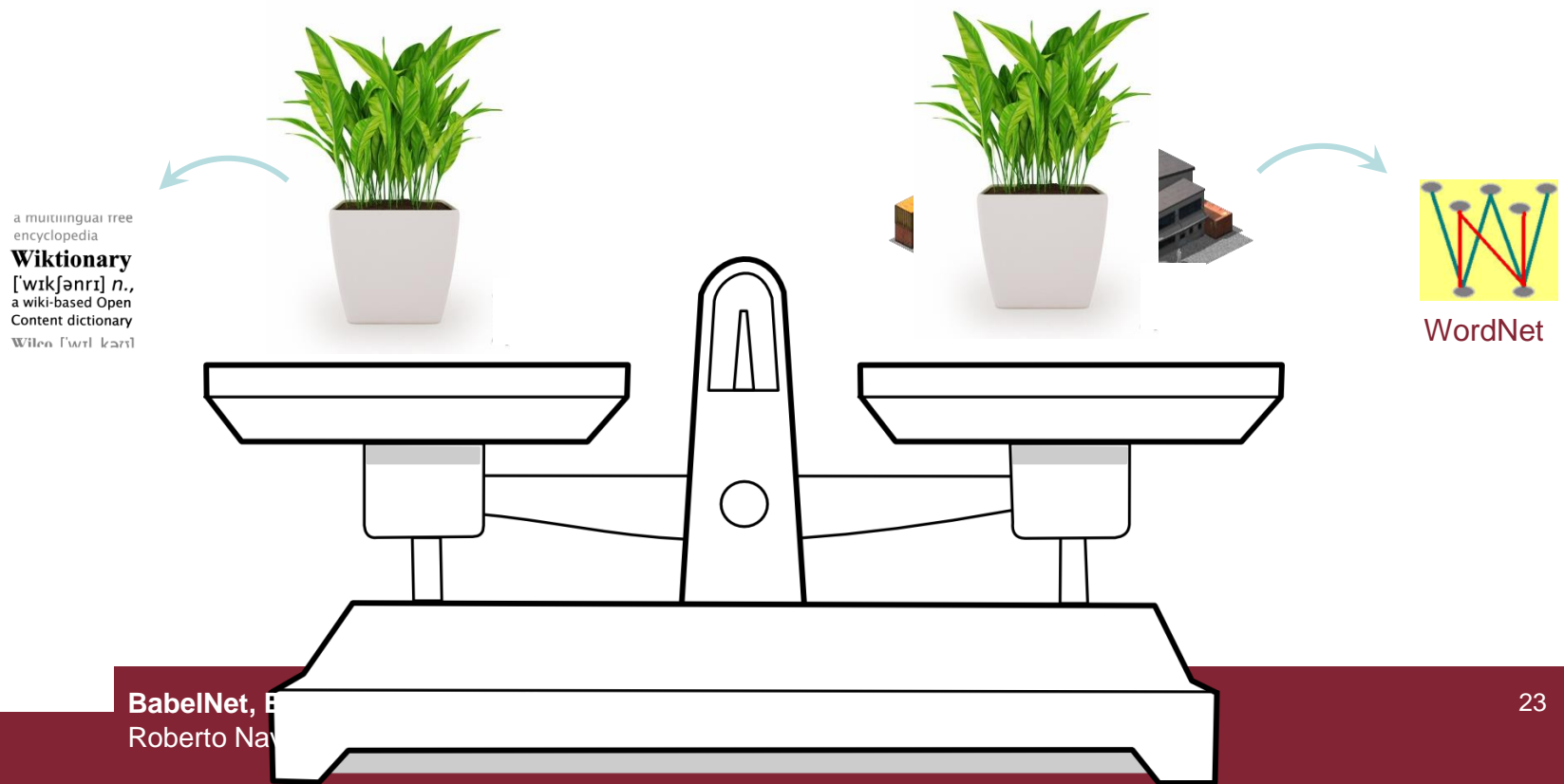


The resulting probabilities form a semantic signature

# To merge or not to merge?

## [Pilehvar and Navigli, ACL 2014]

- Measure the similarity of senses of the same word (but from different resources)
- If they are similar enough, **merge** the corresponding two concepts



# Merging entries from different resources into BabelNet

• **W** **n** **Plant**

Old English *plante* ("young tree or shrub, herb newly planted"), from Latin *planta* ("sprout, shoot, cutting"). Broader sense of "any vegetable life, vegetation generally" is from French *plante*.

The verb is from Middle English *planter*, from Old English *plantian* ("to plant"), from Latin *plantare*, later influenced by Old French *planter*. Compare also Dutch *planter* ("to plant"), German *pflanzen* ("to plant"), Swedish *planter* ("to plant"), Icelandic *planta* ("to plant").

**Pronunciation** [edit]

- (New Zealand, Received Pronunciation) IPA<sup>(key)</sup>: /plɑːnt/
- (Australia, US, Canada, Northern England) IPA<sup>(key)</sup>: /plænt/
- Audio (US) 0:00 ▶ ⏮ ⏭ MENU
- Rhymes: -ɑːnt

**Noun** [edit]

**plant** (plural **plants**)

1. An **organism** that is not an animal, especially an organism capable of photosynthesis. Typically a small or herbaceous organism of this kind, rather than a **tree**. [quotations ▼]  
*The garden had a couple of trees, and a cluster of colourful **plants** around the border.*
2. (**botany**) An **organism** of the kingdom *Plantae*; now specifically, a living organism of the *Embryophyta* (land plants) or of the *Chlorophyta* (green algae), a **eukaryote** that includes double-membraned **chloroplasts** in its cells containing **chlorophyll** *a* and *b*, or any organism closely related to such an organism.

**Definition**

**S: (n) plant, flora, plant life** ((botany) a living organism lacking the power of locomotion)

**S: (n) organism, being** (a living thing that has (or can develop) the power to act or function independently)

- **S: (n) living thing, animate thing** (a living (or once living) organism)
- **S: (n) whole, unit** (an assemblage of parts that is regarded as a single entity) "*how big is that part compared to the whole*" "*the team is a unit*"
- **S: (n) object, physical object** (a tangible and observable entity; an entity that can cast a shadow) "*it was a collection of rackets, balls and other objects*"
- **S: (n) physical entity** (an entity that has physical existence)
- **S: (n) entity** (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Article Talk Read View source View history S

## Plant

From Wikipedia, the free encyclopedia

*For other uses, see **Plant** (disambiguation).*

**Plants**, also called **green plants** (**Viridiplantae** in Latin), are multicellular eukaryotes of the kingdom **Plantae**. They form a clade that includes the flowering plants, conifers and other gymnosperms, ferns, clubmosses, hornworts, liverworts, mosses and the green algae. Plants exclude the red and brown algae, the fungi, archaea, bacteria and animals.

Green plants have cell walls with cellulose and characteristically obtain most of their energy from sunlight via photosynthesis by primary chloroplasts, derived from endosymbiosis with cyanobacteria. Their chloroplasts contain chlorophylls *a* and *b* which gives them their green color. Some plants are parasitic and have lost the ability to produce normal amounts of chlorophyll or to photosynthesize. Plants are also characterized by sexual reproduction, modular and indeterminate growth, and an alternation of generations, although asexual reproduction is also common.

Plants are difficult to determine, but as of 2010, there are thought to be 300–315 thousand species of plants, of which, at majority, some 260–290 thousand, are seed plants (see the table below).<sup>[1]</sup> Green plants provide

**BabelNet**

French: Organisme vivant qui synthétise sa nourriture à partir de substances inorganiques, possède des membranes cellulaires en cellulose, et n'a pas de moyen de locomotion.

1. Ogni organismo vivente che sintetizza il suo cibo da sostanze inorganiche, possiede membrane cellulari di cellulosa, risponde ad uno stimolo, manca di organi di senso specializzati e del sistema nervoso, e non ha poteri di locomozione.

1. Organizm o komórkach pokrytych ścianą komórkową, samożywny, wytwarzający dzięki ciałkom zieleni w drodze asymilacji złożonych związków nieorganicznych lub cudzożywny, odżywiający się związkami organicznymi.

nian: Vsaak živ organism, ki si hrano sintetizira iz anorganskih snovi, proizvajajo celulozne celične stene, se počasi in pogosto trajno specializirani čutilni organi in živčni sistem in se ne more premikati.

ish: En levande organism som typiskt syntetiserar sin föda från oorganiska ämnen, som har cellväggar av cellulosa, som reagerar på stimuli, som saknar specialiserade sinnesorgan och nervsystem, och som saknar rörelseförmåga.

**synonyms and translations**

Language	Spelling	Annotation
Arabic	landare	show ▼
Armenian	plante	show ▼
Basque	plant	show ▼
Bulgarian	растение	show ▼
Catalan	planta	show ▼
Czech	rostlina	show ▼
Dutch	planter	show ▼
English	plant	show ▼
Esperanto	flora	show ▼

**OMEGAWIKI**

## BabelNet: concepts and semantic relations (2)

- We encode knowledge as a **labeled directed graph**:
  - Each vertex is a **Babel synset**



- Each edge is a **semantic relation** between synsets:
  - **is-a** (balloon is-a aircraft)
  - **part-of** (gasbag part-of balloon)
  - **instance-of** (Einstein instance-of physicist)
  - ...
  - **unspecified/relatedness** (balloon related-to flight)

# 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

Ænglisc

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

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

*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.

Statistical Machine Translation

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

# The most frequent translation of a word in a given meaning

left context	term	right context
	wikification	may refer to: the...
geoinformatics services' and '	wikification	of GIS by the masses'
the process may be called	wikification	(as in ...
which is then called "	wikification	and to the related problem
reason needs copyediting,	wikification	, reduction of POV, work on references
huge amount of cleanup,	wikification	, etc. Version of 12 Nov

# The most frequent translation of a word in a given meaning

left context	term	right context
	wikificazione	potrebbe riferirsi a: il...
servizi geoinformatici' e '	wikification	di GIS dalle masse'
il processo chiamato	wikificazione	(come in ...
che è quindi chiamato	wikificazione	e al problema correlato...
ragione richiede copyediting,	wikification	, riduzione di POV, lavoro su reference
grandi quantità di pulizia,	wikificazione	, ecc. Versione del 12 Novembre

# The most frequent translation of a word in a given meaning

left context	term	right context
	wikificazione	potrebbe riferirsi a: il...
servizi geoinformatici' e '	wikification	di GIS dalle masse'
il processo chiamato	wikificazione	(come in ...
che è quindi chiamato	wikificazione	e al problema correlato...
ragione richiede copyediting,	wikification	, riduzione di POV, lavoro su reference
grandi quantità di pulizia,	wikificazione	, ecc. Versione del 12 Novembre

# What is BabelNet?

- A **merger** of resources of different kinds:

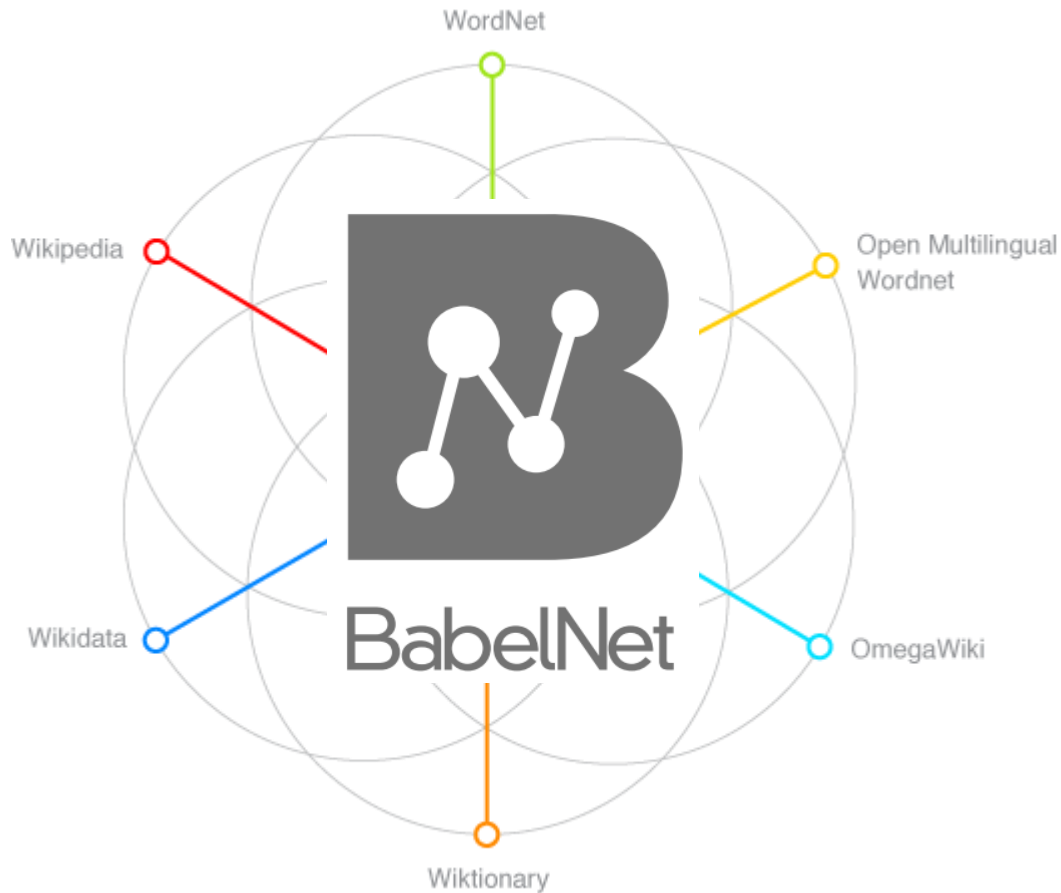


# What is BabelNet?

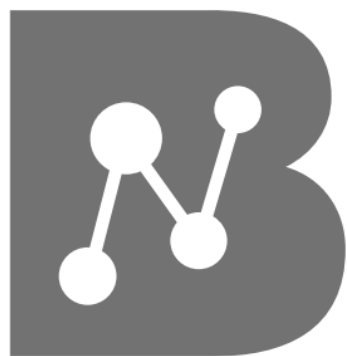
- A **merger** of resources of different kinds:
  - **WordNet**: the most popular computational lexicon of English
  - **Open Multilingual WordNet**: a collection of open wordnets
  - **Wikipedia**: the largest collaborative encyclopedia
  - **Wikidata**: the largest collaborative knowledge base
  - **Wiktionary**: the largest collaborative dictionary
  - **OmegaWiki**: a medium-size collaborative multilingual dictionary
  - High-quality automatic **sense-based translations**

# What is BabelNet?

- A **merger** of resources of different kinds:



# Not to be confused with:



BabelNet

≠



Unbabel


# Why do we need BabelNet?

- **Multilinguality**: the same concept is expressed in tens of languages

- Dictionary
- Images
- Translations
- Sources
- Categories
- External links

English Arabic Chinese French German Greek Hebrew Hindi Italian Japanese + all preferred languages







bn:00002838n • NOUN • Concept • Categories: Bicycle tools, Mechanical hand tools, Screws


 **Allen wrench** • Hex key

A wrench for Allen screws + More definitions

IS-A: wrench • tool • hand tool

EXPLORE NETWORK





# Why do we need BabelNet?


- **Multilinguality**: the same concept is expressed in tens of languages

## Translations

-  مفك سداسي, مفتاح سداسي, آلن وجع, وجع آلين, مفتاح آلين, مفتاح عرافة
-  内六角扳手, 六角匙, 内六角扳手, 内六角扳手, 六角扳手
-  Allen wrench, Hex key, Allen key, Hex head wrench, Allen bolt, Allan keys, Inbus, Alum key, Allan wrench, Zeta key, Allen socket, Hex wrench, Allum key, Unbrako, Alan wrench, Alan key, Allen keys, Imbus, Hex driver, Allan key, Socket head, Umbrako
-  Clé Allen, clef Allen, Clef six pans, Clé six pans creux, *clé hexagonale*
-  Inbusschlüssel, Innensechskantschlüssel, Innensechskant, Inbus, Inbusschraube, Innensechskantschraube Bauer und Schaurte, Sechskantschlüssel, Innensechskantschraube, Sechskantschraubendreher
-  κλειδί allen, εξαγωνα κλειδί
-  מפתח אלן, אלן מפתח ברגים, מפתח ברגים, אלן מפתח
-  एलन रॉच, हेक्स कुंजी
-  Chiave a brugola, Brugola, Viti brugola, Imbus, Chiave di Allen, Chiave Allen, *chiave esagonale*
-  六角棒スパナ, 六角レンチ, 六角棒レンチ, ヘキサゴンレンチ, アーレンキー, 六角レンチ。
-  Шестигранный ключ, Шестигранный шлиц, Инбусовый ключ, Инбус, Имбусовый ключ, Шестигранник
-  llave allen, Llaves allen, Tornillo allen, *llave hexagonal*

# Why do we need BabelNet?

- **Multilinguality**: the same concept is expressed in tens of languages



BabelNet

LOG IN REGISTER


allen wrench ENGLISH TRANSLATE INTO... SEARCH

PREFERENCES


English Arabic Chinese French German Greek Hebrew Hindi Italian Japanese + all preferred languages

Dictionary  
Images  
Translations  
Sources  
Categories  
External links


bn:00002838n • NOUN • Concept • Categories: Bicycle tools, Mechanical hand tools, Screws Categories: براغي, آلات, تقنية Categories: Attrezzi per meccanica

 **Allen wrench** • **Hex key**

A wrench for Allen screws  
+ More definitions

 **مفك سداسي**

مفك سداسي أو مفك سداسي الأضلاع أو مفتاح سداسي أو مفتاح سداسي الأضلاع هو أداة ذات مقطع عرضي سداسي الأضلاع لفك البراغي.

 **Brugola**

Una chiave a brugola o brugola, denominata più correttamente chiave di Allen ma conosciuta anche in gergo tecnico

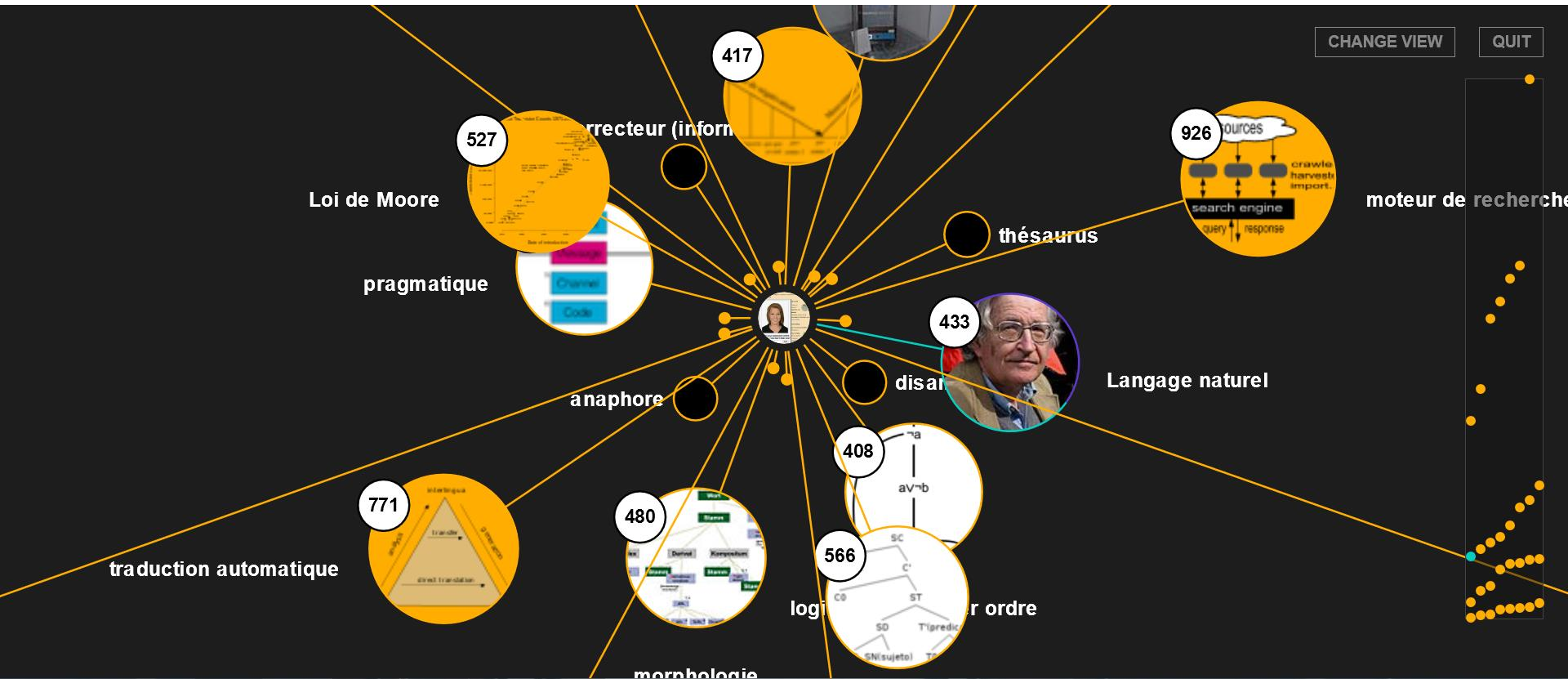
# Why do we need BabelNet?

- **Multilinguality**: the same concept is expressed in tens of languages
- **Coverage**: 271 languages and 14 million entries!



## Why do we need BabelNet?

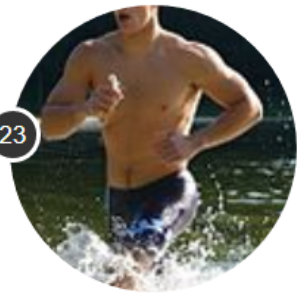
- **Multilinguality**: the same concept is expressed in tens of languages
- **Coverage**: 271 languages and 14 million entries!
- **Concepts and named entities together**: dictionary and encyclopedic knowledge is semantically interconnected



# Why do we need BabelNet?

- **Multilinguality**: the same concept is expressed in tens of languages
- **Coverage**: 271 languages and 14 million entries!
- **Concepts and named entities together**: dictionary and encyclopedic knowledge is semantically interconnected
- **"Dictionary of the future"**: semantic network structure with labeled relations, pictures, multilingual synsets

Verb



run

Move fast by using one's feet, with one foot off the ground at any given time

ID: 00093170v | Concept

هَرُؤَل, جَرَى, رَغَضَن

奔跑, 跑

courir

rennen

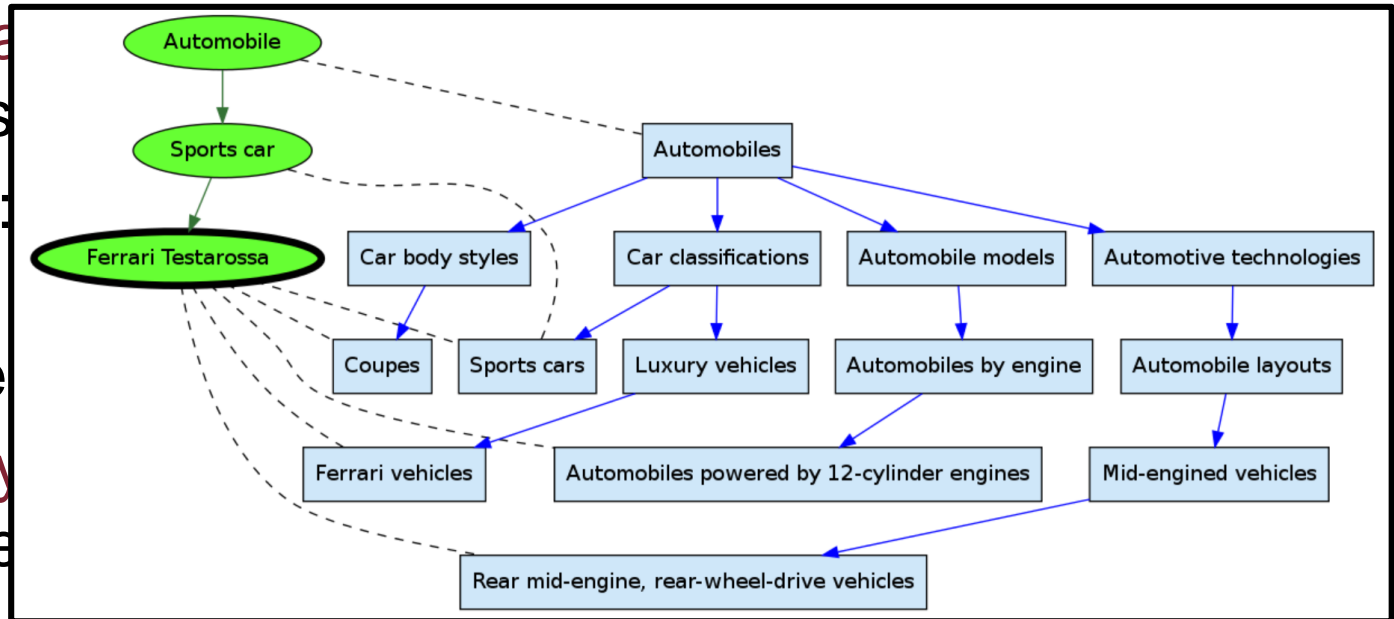
τρέχω, κινούμαι

ץר

correre

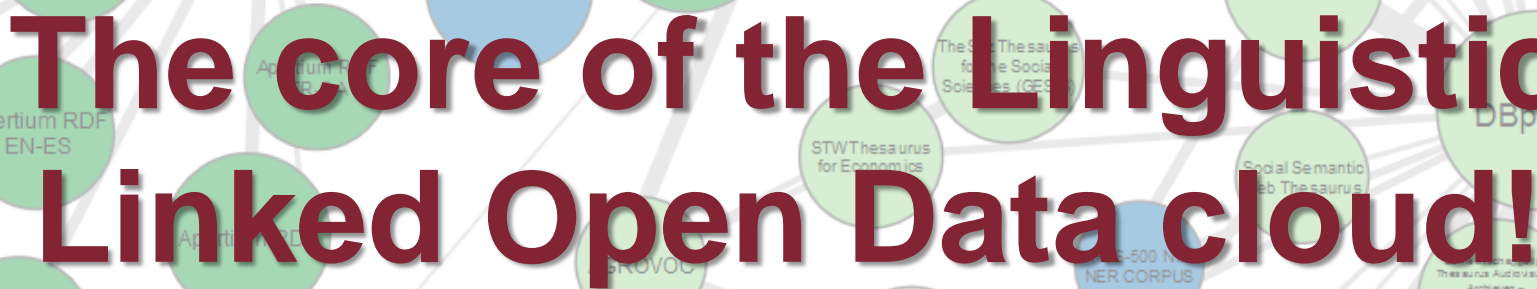
# Why do we need BabelNet?

- Multilingual languages
- Coverage
- Concepts encyclopedic
- "Dictionary with labels"
- **Full-fledged taxonomy:** is-a relations are available for both concepts and named entities (**Wikipedia Bitaxonomy**)
  - Lisbon *is-a* city & port & capital & provincial capital & national capital
  - BabelNet *is-a* semantic network & encyclopedic dictionary
  - summer school *is-a* academic term
  - Ferrari Testarossa *is-a* sports car



# Why do we need BabelNet?

- **Multilinguality**: the same concept is expressed in tens of languages
- **Coverage**: 271 languages and 14 million entries!
- **Concepts and named entities together**: dictionary and encyclopedic knowledge is semantically interconnected
- **"Dictionary of the future"**: semantic network structure with labeled relations, pictures, multilingual synsets
- **Full-fledged taxonomy**: is-a relations are available for both concepts and named entities (**Wikipedia Bitaxonomy**)
- **Easy access**: Java and HTTP RESTful APIs; SPARQL endpoint (2 billion triples)



# What can we do with BabelNet?

- Search and translate:

The screenshot displays the BabelNet website. At the top center is the BabelNet logo, a stylized 'B' with a network diagram inside, and the text 'BabelNet' below it. A tagline reads 'A very large multilingual encyclopedic dictionary and semantic network'. In the top right corner, there are links for 'LOG IN' and 'REGISTER'. The main interface features a search bar with the word 'plane' entered. To the right of the search bar are dropdown menus for 'ENGLISH' and '3 SELECTED'. A teal 'TRANSLATE' button is on the far right. Below the search bar, a teal banner states 'THE BABELNET 3.0 JAVA & HTTP APIS ARE AVAILABLE NOW'. A 'PREFERENCES' link with a gear icon is visible. A language selection dropdown is open, showing a list of languages with checkboxes: ARABIC, CHINESE (checked), FRENCH, GERMAN (checked), GREEK, HEBREW, HINDI, and ITALIAN (checked). The footer contains a small BabelNet logo, a list of links (ABOUT, PUBLICATIONS, STATS, DOWNLOADS, API GUIDE), a paragraph of text about the project's funding and license, and logos for 'STUDIVM PARIS' and 'erc'.

plane

ENGLISH

3 SELECTED

TRANSLATE

Search

THE BABELNET 3.0 JAVA & HTTP APIS ARE AVAILABLE NOW

⚙️ PREFERENCES

☐ ARABIC

☒ CHINESE

☐ FRENCH

☒ GERMAN

☐ GREEK

☐ HEBREW

☐ HINDI

☒ ITALIAN

ABOUT  
PUBLICATIONS  
STATS  
DOWNLOADS  
API GUIDE

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

STUDIVM PARIS

erc

# What can we do with BabelNet?

- Noun
- Verb
- Adjective

## Noun



### airplane, plane, aeroplane

An aircraft that has a fixed wing and is powered by propellers or jets

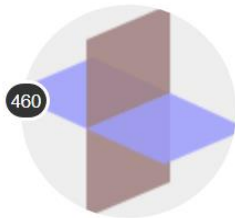
ID: 00001697n | Concept

固定翼飛機, 飛行機, 飞龙机

avion, aéroplane

Flugzeug

aereo, aeroplano, apparecchio



### plane, sheet

(mathematics) an unbounded two-dimensional shape

ID: 00062766n | Concept

平面, 面

plan

Ebene (Mathematik)

piano, piano geometrico



### plane

A level of existence or development

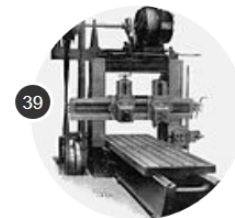
ID: 00062767n | Concept

平面的存在

plan

Ebene

piano, Spostamento della realtà, livello



### planer, plane, planing machine

A power tool for smoothing or shaping wood

ID: 00062768n | Concept

刨床

raboteuse, rabot

Hobelmaschine

piallatrice



### plane, woodworking plane, carpenter's plane

A carpenter's hand tool with an adjustable blade for smoothing or shaping wood

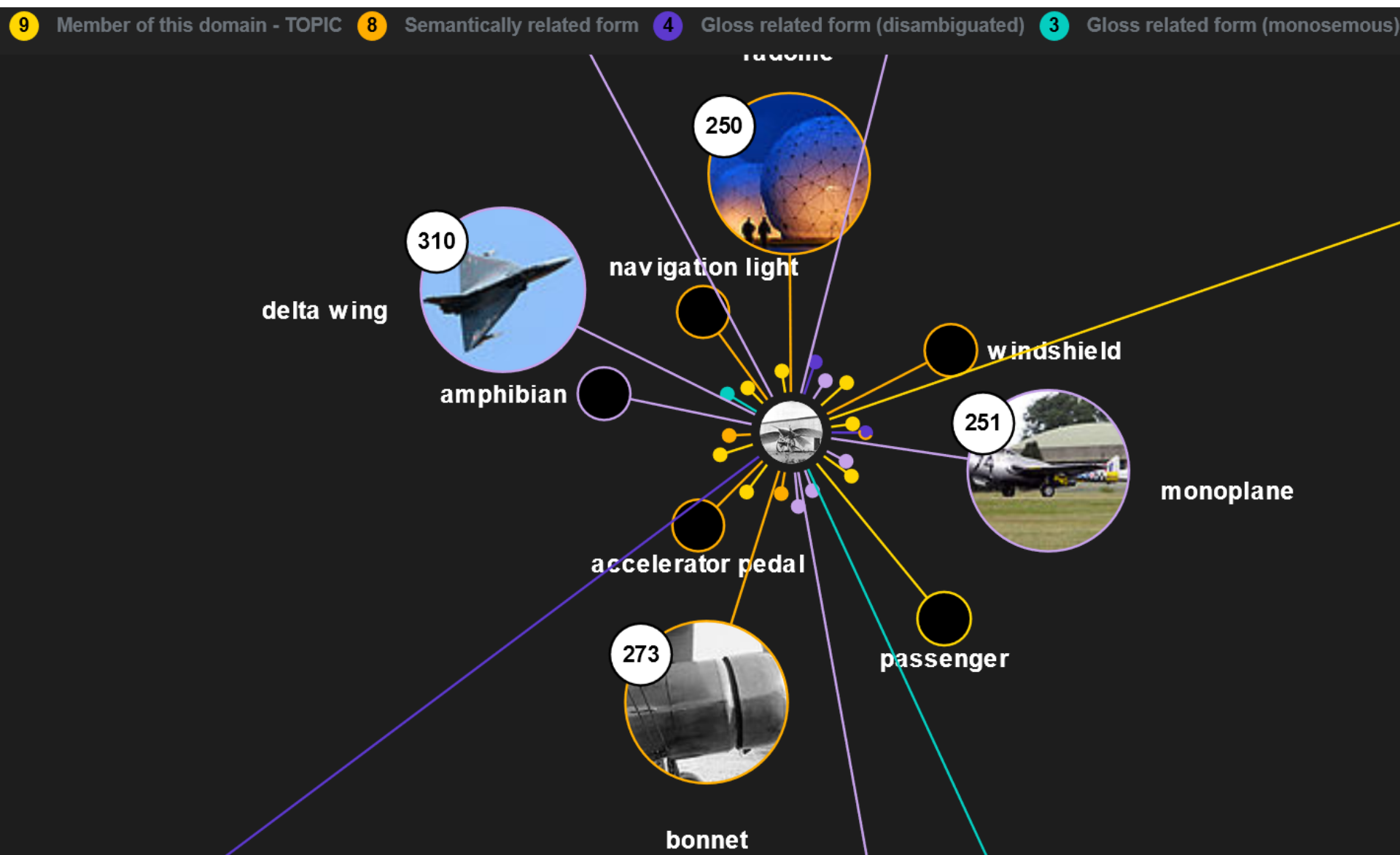
刨


rabot, avion, appareil

Hobel

# What can we do with BabelNet?

- Explore the network:



A full-page photograph of a night sky filled with stars and the Milky Way galaxy. In the lower right foreground, the dark silhouette of a person stands looking up at the stars. The sky is a deep black, densely populated with white stars of varying sizes. The Milky Way's luminous band stretches diagonally across the frame from the upper left towards the lower right. The overall mood is contemplative and awe-inspiring.

*“Interestingly, the feeling of being all alone  
in the entire Universe can be  
mystically beautiful”*

# We are not alone in the (resource) universe!

- **DBpedia** [Bizer et al. 2009] - a resource obtained from structured information in **Wikipedia**
  - «Describes 3.77M things»
  - No dictionary side
- **YAGO** [Suchanek et al. 2007]
  - «Contains 10M entities and 120M facts about these entities»
  - Links **Wikipedia** categories to **WordNet** synsets (**most freq. sense**)
- **MENTA** [de Melo and Weikum, 2010]
  - A «multilingual taxonomy with 5.4M entities»
- **WikiNet** [Nastase and Strube, 2013]
  - Semantic network connecting **Wikipedia** entities
  - «3M concepts and 38+M relations»
- **Freebase** (<http://freebase.com>): collaborative effort
  - Started from **Wikipedia**, MusicBrainz, ChefMoz, etc. **Shut down!**

# PREVIEW: BabelNet 3.1 will be a knowledge base!

- Wikidata + Infoboxes (superset of DBpedia) + relations extracted with Open Information Extraction techniques + domain labels

● Dictionary

● Images

● Translations

● Sources

● Categories

● External links

 **Steve Jobs** · **Steve Jobs**

Steven Paul Jobs, dit Steve Jobs, est un entrepreneur et inventeur américain, souvent qualifié de visionnaire, et une figure majeure de l'électronique grand public, notamment pionnier de l'avènement de l'ordinateur personnel, du baladeur numérique, du smartphone et de la tablette tactile. [+ More definitions](#)

IS-A: [homme](#)

— [Less relations](#)

RELATED:

 [Inside Apple](#)

 [NeXT Introduction](#)

 [The Little Kingdom](#)

[VENUS YACHT](#)

[Bill Fernandez](#)

[Rob Janoff](#)

[ISteve](#)

[Burrell Smith](#)

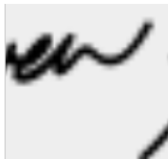
[Daniel Kottke](#)

[Bud Tribble](#)

 [NeXT MegaPixel ...](#)

CHILD: [Lisa Brennan-Jobs](#) AWARD-RECEIVED: [National Medal of Technology and Innovation](#) SEX-OR-GENDER: [masculin](#) GIVEN-NAME: [Steven](#)  
CAUSE-OF-DEATH: [cancer du pancréas](#) SPOUSE: [Laurene Powell Jobs](#) OCCUPATION: [Ingénieur](#) · [Entrepreneur](#) · [inventeur](#) · [cadre supérieur](#) ·  
[Inventeur](#) EMPLOYER: [Apple](#) EDUCATED-AT:  [Homestead High School \(Cupertino, California\)](#) · [Reed College](#) COUNTRY: [États-Unis](#) RELIGION:  
[Bouddhisme](#) PLACE-OF-BIRTH: [San Francisco](#) PLACE-OF-DEATH: [Palo Alto](#) STATED-IN: [Gemeinsame Normdatei](#) SISTER: [Mona Simpson \(auteur\)](#)  
LOCATED-IN-THE-ADMINISTRATIVE-TERRITORIAL-ENTITY: [Comté de Santa Clara](#) · [Californie](#)

EXPLORE NETWORK



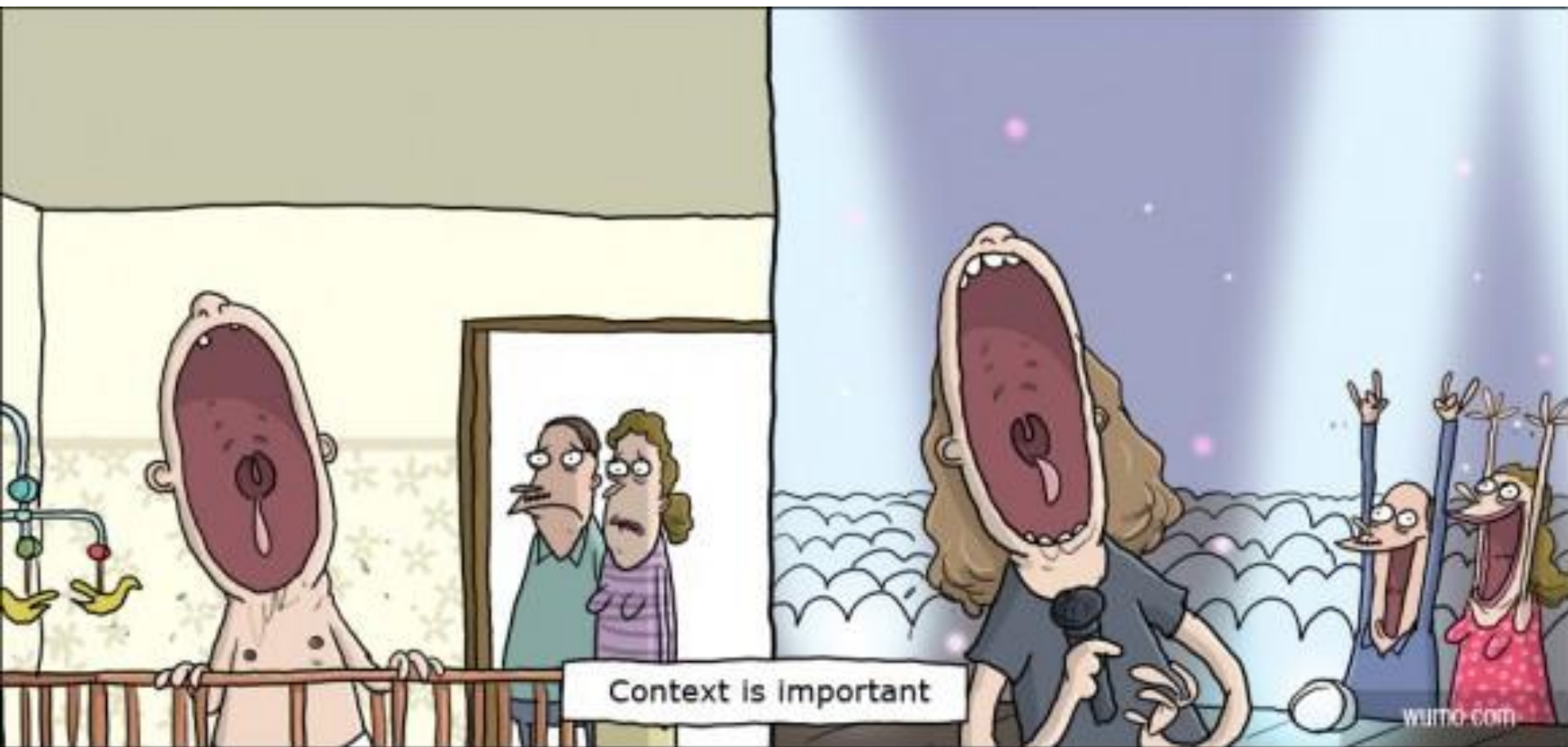
**ONE DOES NOT SIMPLY USE BABELNET**

**AS A MULTILINGUAL DICTIONARY**

# ADDRESSING AMBIGUITY

[Moro, Raganato & Navigli,  
TACL 2014]

# CONTEXT MATTERS!!!



# More seriously: lexical ambiguity!

- Thomas and Mario played as strikers in Munich.

Thomas

and

Mario

played

as

strikers

in

Munich



Thomas

Thomas Müller is a German footballer who plays for Bayern Munich and the



Mario

Mario Gómez García is a German footballer who plays as a striker for Bayern Munich in

played

participate in games or sport; "We played hockey all afternoon"; "play cards"; "Pele



strikers

a forward on a soccer team



Munich

FC Bayern Munich, is a German sports club based in Munich, Bavaria.

# Word Sense Disambiguation and Entity Linking

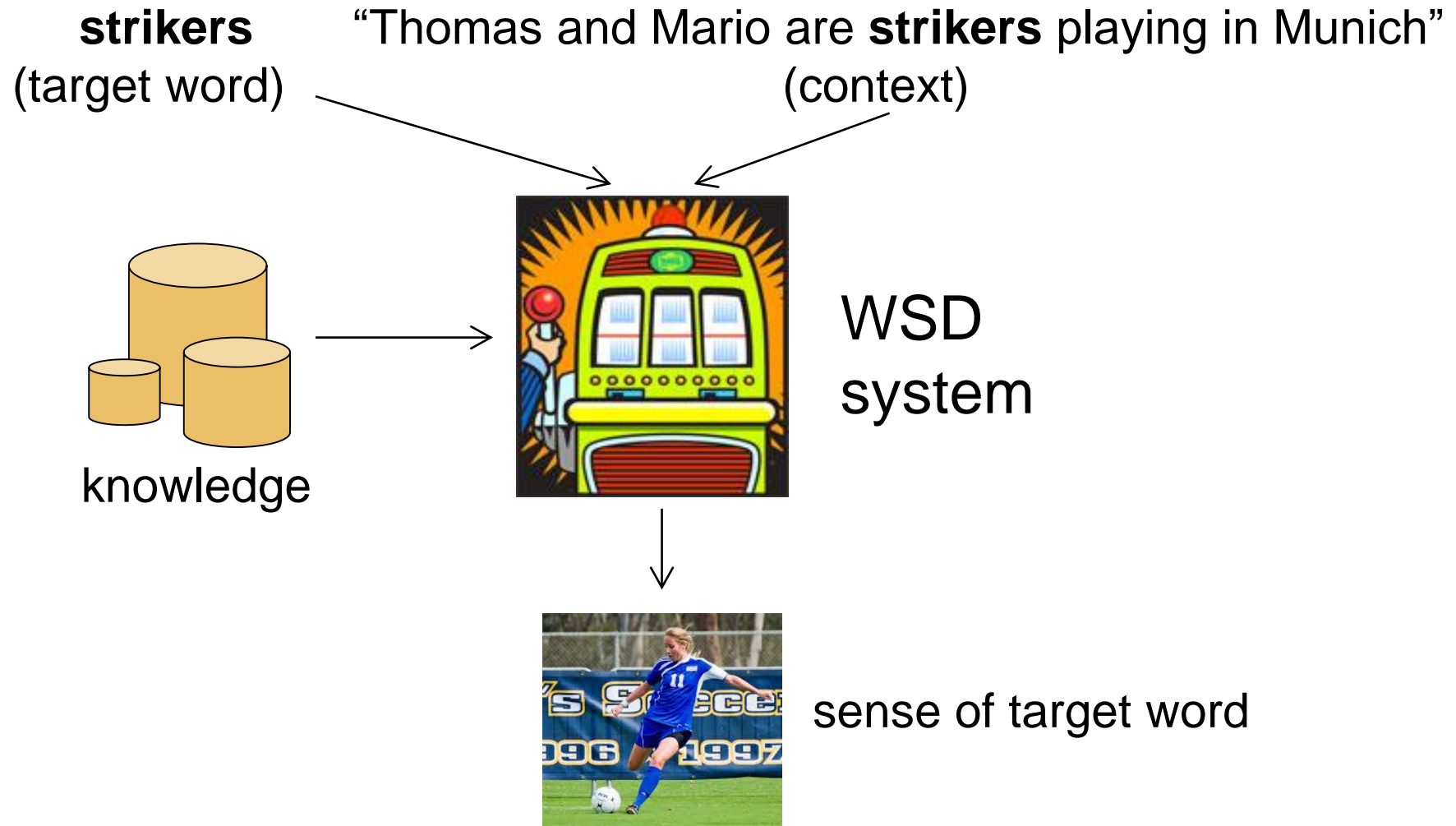
- Thomas and Mario are strikers playing in Munich



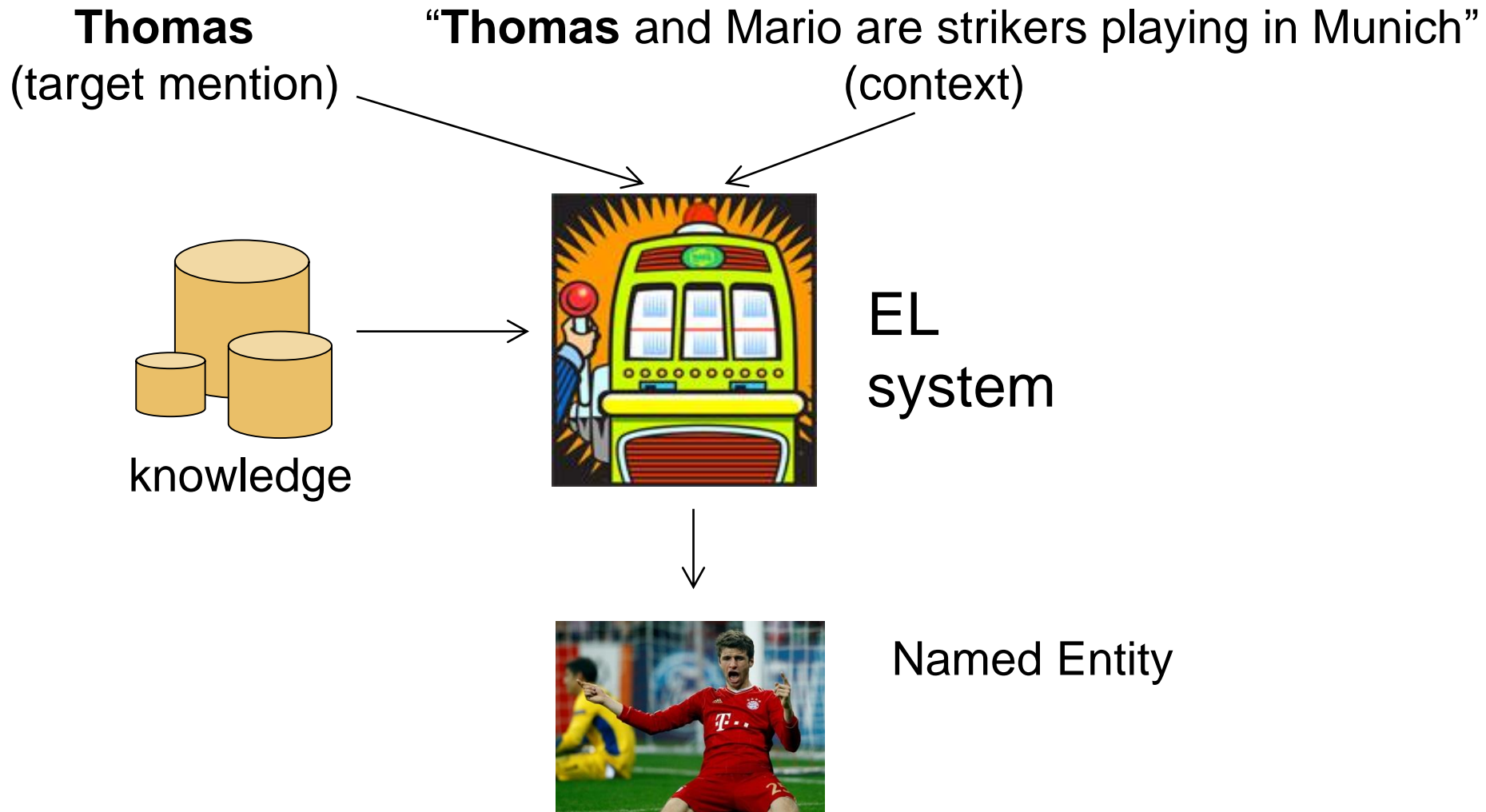
**Entity Linking:** The task of discovering mentions of entities within a text and linking them in a knowledge base.

**WSD:** The task aimed at assigning meanings to word occurrences within text.

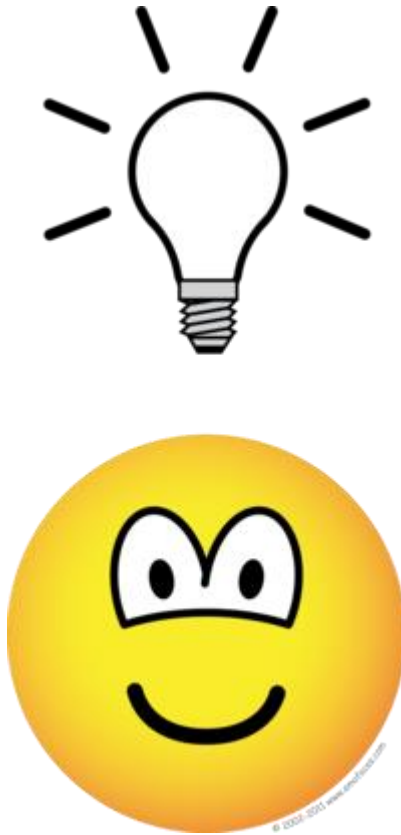
# Word Sense Disambiguation in a Nutshell



# Entity Linking in a Nutshell



# Disambiguation and Entity Linking together!



BabelNet is a huge **multilingual inventory**  
for both word senses and named entities!

# Multilingual Joint Word Sense Disambiguation (MultiJEDI)

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



So what?



Babelfy

# Step 1: Find all possible meanings of words

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

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



Thomas,  
Norman



Thomas,  
Seth



They both have  
Thomas as one of  
their lexicalizations

# Step 1: 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

# Step 1: 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



Mario Gómez



Striker (Movie)



Munich (Song)



Thomas (novel)



# Step 1: 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)



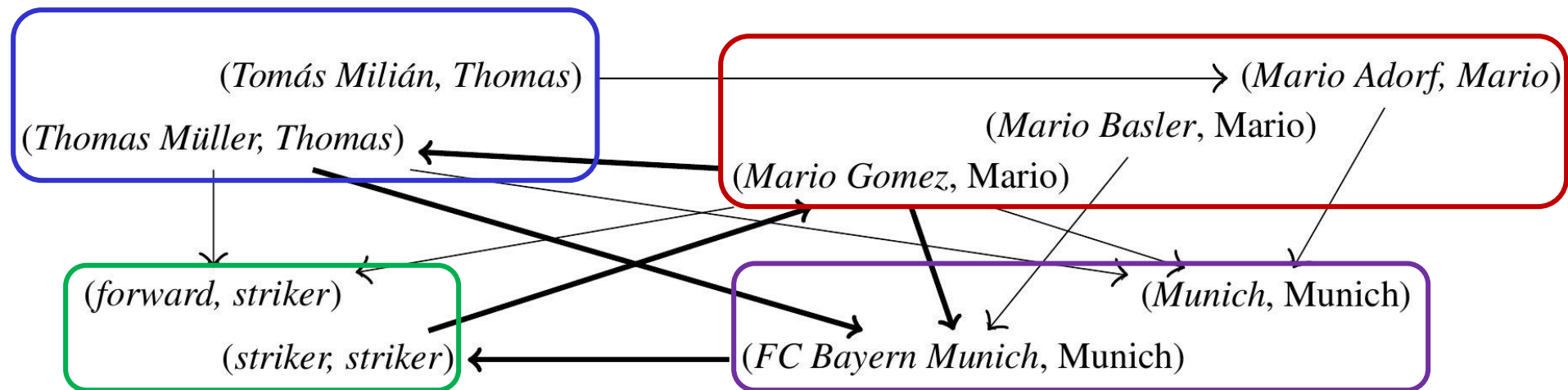
Munich (Song)



**Ambiguity!**

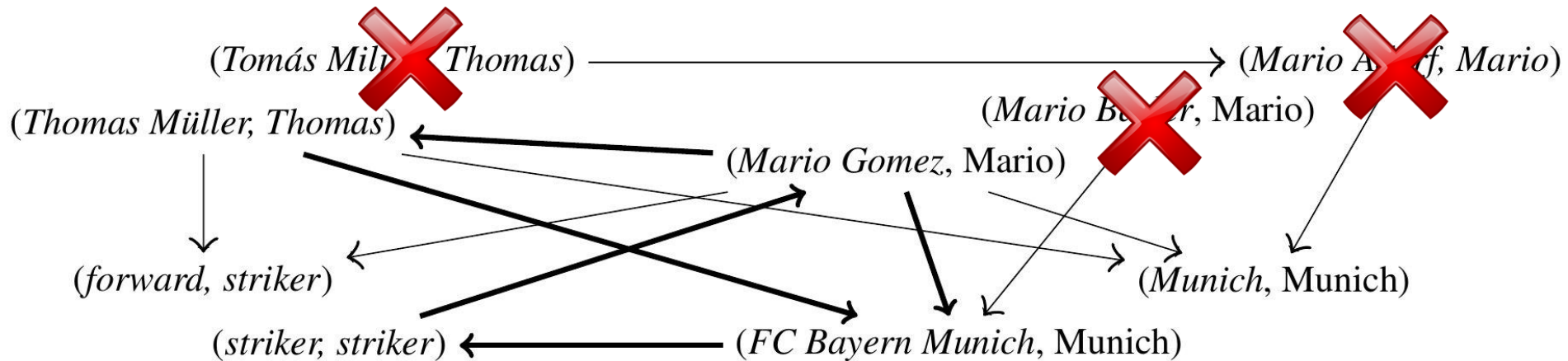
## Step 2: Connect all the candidate meanings

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



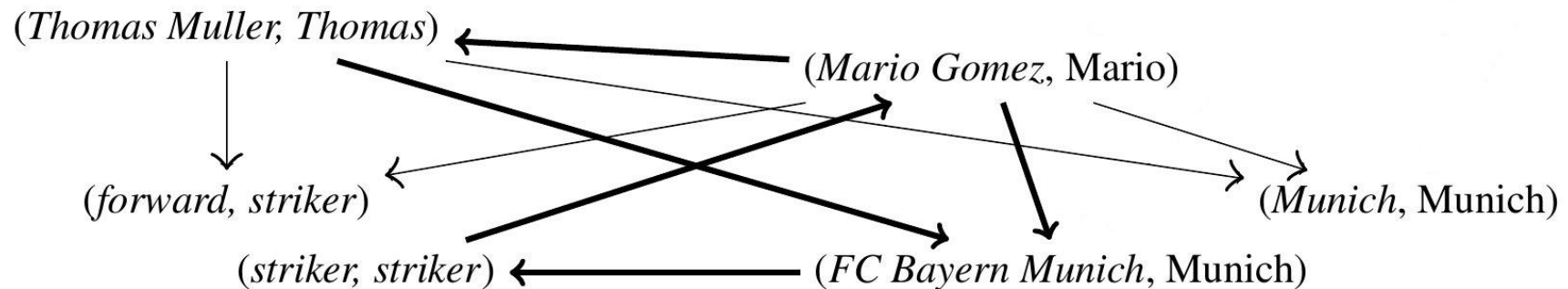
## Step 3: Extract a dense subgraph

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



## Step 3: Extract a dense subgraph

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



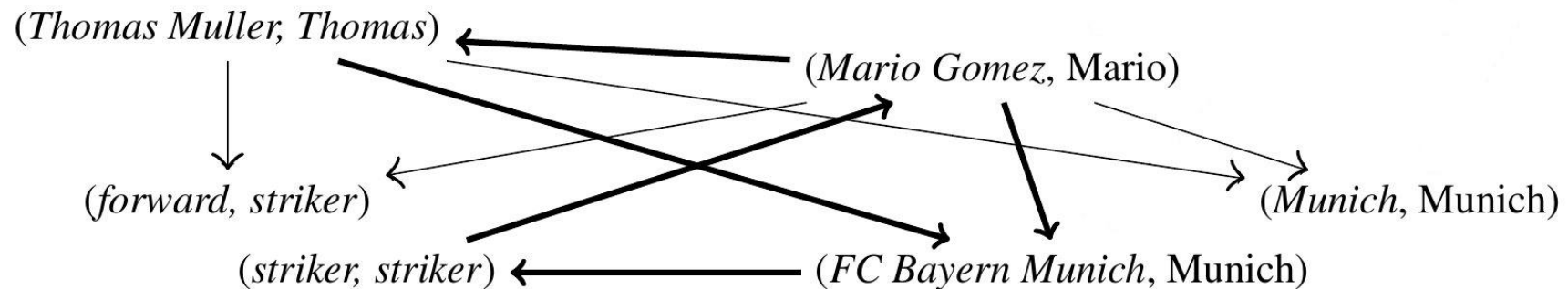
## Step 4: Select the most reliable meanings

$$w_{(v,f)} := \frac{|\{f' \in F : \exists v' \text{ s.t. } ((v, f), (v', f')) \text{ or } ((v', f'), (v, f)) \in E_I\}|}{|F| - 1}$$
$$score((v, f)) = \frac{w_{(v,f)} \cdot deg((v, f))}{\sum_{v' \in cand(f)} w_{(v',f)} \cdot deg((v', f))}$$

- We take into account both the **lexical coherence**, in terms of the number of fragments a candidate relates to, and the **semantic coherence**, using a graph centrality measure among the candidate meanings.

## Step 4: Select the most reliable meanings

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



# Step 4: Select the most reliable meanings

“Thomas and Mario are strikers playing in Munich”

Seth Thomas



Mario (Character)



striker (Sport)



Munich (City)



Thomas Müller



Mario (Album)



Striker (Video Game)



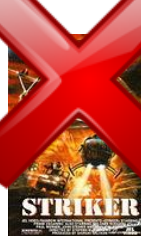
FC Bayern Munich



Mario Gómez



Striker (Movie)



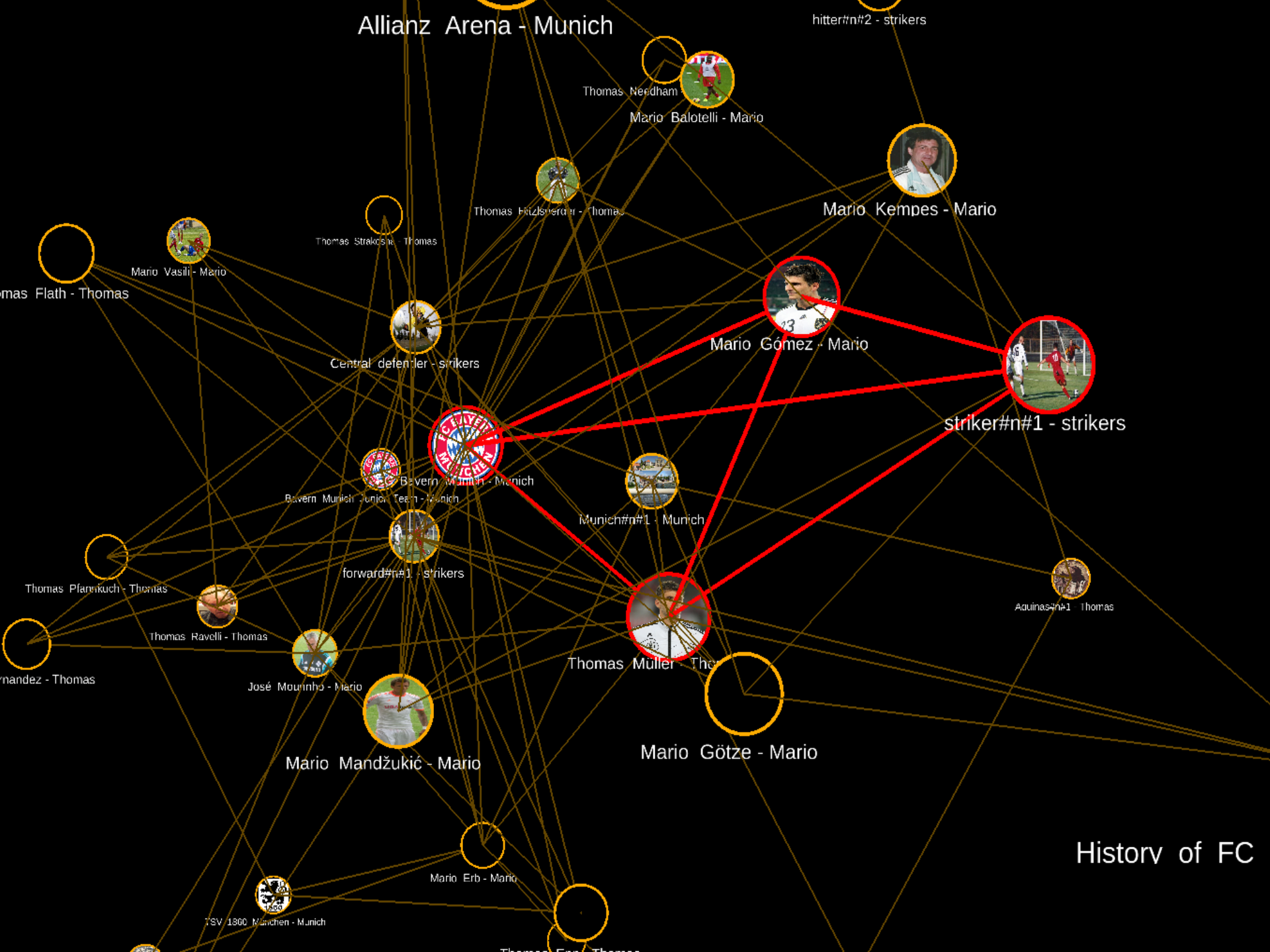
Munich (Song)



Thomas (novel)



# Allianz Arena - Munich



History of FC

# Experimental Results:

## Fine-grained (Multilingual) Disambiguation

Senseval-3      SemEval-2007 task 17      SemEval-2013 task 12

	Sens3	Sem07	SemEval-2013 English			French		German		Italian		Spanish	
System	WN	WN	WN	Wiki	BN	Wiki	BN	Wiki	BN	Wiki	BN	Wiki	BN
Babelfy	68.3	62.7	<b>65.9</b>	<b>87.4</b>	<b>69.2</b>	<b>71.6</b>	*56.9	81.6	<b>69.4</b>	<b>84.3</b>	66.6	<b>83.8</b>	69.5
IMS	<b>71.2</b>	63.3	65.7	–	–	–	–	–	–	–	–	–	–
UKB w2w	*65.3	*56.0	61.3	–	60.8	–	<b>60.8</b>	–	66.2	–	<b>67.3</b>	–	70.0
UMCC-DLSI	–	–	64.7	54.8	68.5	*60.5	60.5	*58.1	62.8	*58.3	65.8	*61.0	<b>71.0</b>
DAEBAK!	–	–	–	–	60.4	–	53.8	–	59.1	–	*61.3	–	60.0
GETALP-BN	–	–	51.4	–	58.3	–	48.3	–	52.3	–	52.8	–	57.8
MFS	70.3	<b>65.8</b>	*63.0	*80.3	*66.5	69.4	45.3	<b>83.1</b>	*67.4	82.3	57.5	82.4	*64.4
Babelfy unif. weights	67.0	65.2	65.0	87.0	68.5	71.9	57.2	81.2	69.8	83.7	66.8	83.8	70.8
Babelfy w/o dens. sub.	68.3	63.3	65.4	87.3	68.7	71.6	57.0	81.7	69.1	84.4	66.5	83.9	69.5
Babelfy only concepts	68.2	62.7	65.5	83.0	68.7	70.2	56.6	79.3	69.3	83.0	66.3	84.0	69.7
Babelfy on sentences	66.0	65.2	63.5	84.0	67.1	70.7	53.6	82.3	68.1	83.8	64.2	83.5	68.7

# Experimental Results: KORE50, AIDA-CoNLL

- Two gold-standard Entity Linking datasets:

System	KORE50	CoNLL
Babelfy	<b>71.5</b>	82.1
KORE-LSH-G	64.6	81.8
KORE	63.9	*80.7
MW	*57.6	<b>82.3</b>
Tagme	56.3	70.1
KPCS	55.6	82.2
KORE-LSH-F	53.2	81.2
UKB w2w (on BabelNet)	52.1	71.8
Illinois Wikifier	41.7	72.4
DBpedia Spotlight	35.4	34.0
Babelfy unif. weights	69.4	81.7
Babelfy w/o dens. sub.	62.5	78.1
Babelfy only NE	68.1	78.8

# WSD and Entity Linking together win!

I was so lucky I could drive a Ferrari Testarossa !

lucky

Occurring by chance



drive

Operate or control a vehicle



Ferrari Testarossa

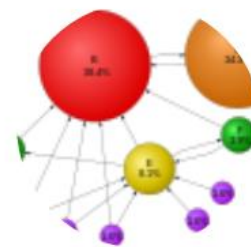
The Ferrari Testarossa is a 12-cylinder

We wrote PageRank in Java .



wrote

Create code, write a computer program



PageRank

PageRank is an algorithm used by Google Search to rank websites in their



Java

A platform-independent object-oriented programming language

# What can we do with Babelfy?

- Disambiguate text written in **any** language!

The screenshot shows the BabelNet website interface. At the top left is the BabelNet logo. To its right is a search bar containing the text "Twitter potenzia chat privata, si può comunicare con tutti." Below the search bar, the word "ITALIAN" is displayed. To the right of the search bar is a "SEARCH" button. In the top right corner, there are links for "LOG IN" and "REGISTER". Below the search bar, there is a "TRANSLATE INTO..." dropdown menu and a "PREFERENCES" link. A light blue banner below the search bar states: "This looks like a sentence: loading Babelfy! Please wait." Below the banner, there are two view options: "expanded view" (selected) and "compact view". The main content area displays the sentence "Twitter potenzia chat privata, si può comunicare con tutti." with each word highlighted in a green bubble. Below each word is a card providing disambiguation information:

- Twitter**: Twitter is an online social networking service that enables users to send and
- potenzia**: Online chat may refer to any kind of communication over the Internet that offers
- chat**: Online chat may refer to any kind of communication over the Internet that offers
- privata**: Confined to particular persons or groups or providing privacy
- , si**: To have the ability to do something
- può**: To have the ability to do something
- comunicare**: Transmit information
- con**: Transmit information

# Live demo

- [http://www.dn.pt/inicio/portugal/interior.aspx?content\\_id=4691087&page=-1](http://www.dn.pt/inicio/portugal/interior.aspx?content_id=4691087&page=-1)
- [http://www.dn.pt/inicio/globo/interior.aspx?content\\_id=4691336](http://www.dn.pt/inicio/globo/interior.aspx?content_id=4691336)

## PGR confirma investigações relacionadas com a PT

por P.J. com Lusa Hoje Comentar



MP estará a investigar o envolvimento político de governantes portugueses e brasileiros no negócio da venda dos 50% da Vivo, detidos pela PT, à Telefónica, e a compra de ações da OI pela PT.

A Procuradoria-Geral da República confirma que "existem investigações em curso relacionadas com a PT, as quais se encontram em segredo de justiça".

A confirmação surge depois de o jornal *Público* ter avançado, na edição de hoje, que o Ministério Público está a investigar o

### FERRAMENTAS



### PARTILHAR NOTÍCIA

f Share 4 Tweet 2

in Share 0 g+1

f Gosto 3

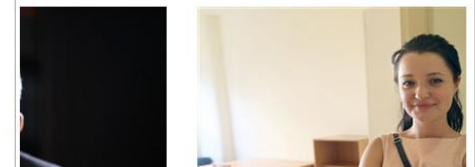
### TAGS

[Portugal](#)



PUB

### NOTÍCIAS MAIS VISTAS



# What can we do with Babelfy?

- Disambiguate in a **language-agnostic** setting!

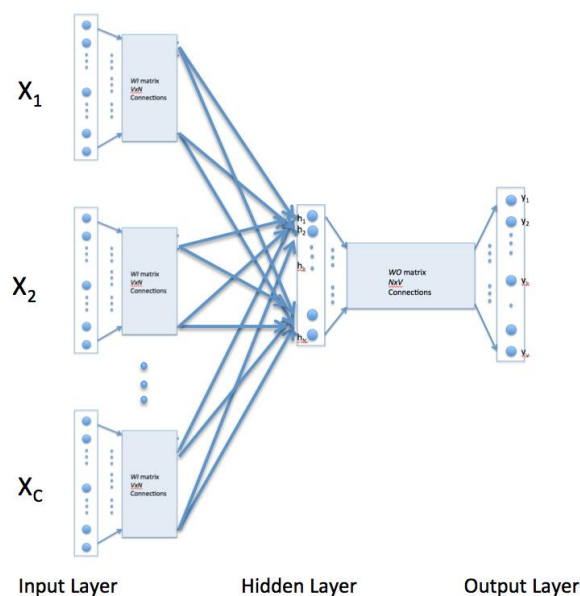
The screenshot displays the Babelfy web interface. At the top left is the Babelfy logo. On the right, there are links for 'LOG IN' and 'REGISTER'. The main input area contains the sentence: 'Technology. Samsung S5 fingerprint flaw exposed. Problema impronte e sblocco telefono galaxy s5.'. Below the input, there is a checkbox for 'Enable partial matches:' which is currently unchecked. To the right of the checkbox is a dropdown menu set to 'AGNOSTIC' and a teal button labeled 'BABELFY!'. Below the input area, there are two tabs: 'expanded view' (selected) and 'compact view'. The 'expanded view' shows a horizontal sequence of terms with corresponding cards below them: 'Technology.' (card: 'Samsung S5 is an Android'), 'Samsung S5' (card: 'fingerprint A print made by an impression of the'), 'fingerprint' (card: 'flaw An imperfection in an object or machine'), 'flaw' (card: 'exposed Remove all or part of one's clothes to show one's body'), 'exposed' (card: 'Problema Defectiveness or unsoundness'), 'Problema' (card: 'impronte A print made by an impression of the'), and 'impronte' (card: 'A print made by an impression of the'). Each card features a circular image related to the term: a Samsung S5 phone, a fingerprint, a flaw on a surface, a person removing clothes, a fingerprint, and a fingerprint.

## Live demo (2) – Crazy polyglot!

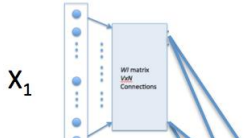
- The 5th Lisbon Machine Learning Summer School
- ha luogo nel mese di luglio 16-23 presso l' Instituto Superior Técnico,
- una escuela líder en Ingeniería y Ciencia en Portugal.
- Elle est organisée conjointement par IST, l'Instituto de Telecomunicações et le Spoken Language Systems Lab – L2F of INESC-ID.
- 點 擊 這 裡 了 解 過 去 的 版 本 信 息

# Sense embeddings [Iacobacci et al., ACL 2015]: explicit meanings and Neural Networks together!!!

- **SensEmbed idea**: moving from latent representation of words to **embeddings of senses**
- **How**: **disambiguate** the **entire English Wikipedia** with Babelfy
- **CBOV**, 5-word window, 400 dimensions, learn **sense embeddings**



# Sense embeddings [Iacobacci et al., ACL 2015]: explicit meanings and Neural Networks together!!!

- **SensEmbed idea**: moving from latent representation of words to **embeddings of senses**
- **How**: **disambiguate** the **entire English Wikipedia** with Babelfy
- **CBOV**, 5-word window, 400 dimensions, learn **sense embeddings**
- Closest senses:  ambiguous words:

$bank_1^n$ (geographical)	$bank_2^n$ (financial)	$number_4^n$ (phone)	$number_3^n$ (acting)	$hood_1^n$ (gang)	$hood_{12}^n$ (convertible car)
upstream $_1^r$	commercial_bank $_1^n$	calls $_1^n$	appearing $_6^v$	tortures $_5^n$	taillights $_1^n$
downstream $_1^r$	financial_institution $_1^n$	dialled $_1^v$	minor_roles $_1^n$	vengeance $_1^n$	grille $_2^n$
runs $_6^v$	national_bank $_1^n$	operator $_{20}^n$	stage_production $_1^n$	badguy $_1^n$	bumper $_2^n$
confluence $_1^n$	trust_company $_1^n$	telephone_network $_1^n$	supporting_roles $_1^n$	brutal $_1^a$	fascia $_2^n$
river $_1^n$	savings_bank $_1^n$	telephony $_1^n$	leading_roles $_1^n$	execution $_1^n$	rear_window $_1^n$
stream $_1^n$	banking $_1^n$	subscriber $_2^n$	stage_shows $_1^n$	murders $_1^n$	headlights $_1^n$

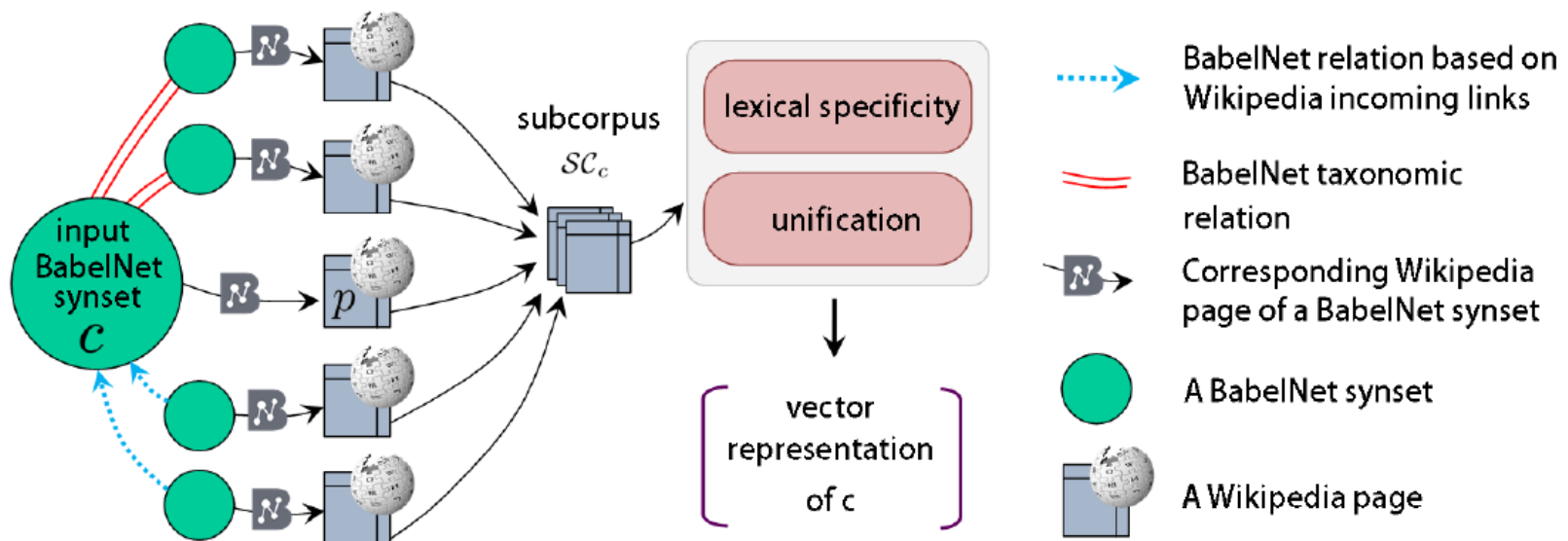
## Sense embeddings [Iacobacci et al., ACL 2015]: explicit meanings and Neural Networks together!!!

- **SensEmbed idea:** moving from latent representation of words to **embeddings of senses**
- State-of-the-art performance **beyond word2vec**:

Measure	Dataset					Average
	RG-65	WS-Sim	WS-Rel	YP-130	MEN	
Pilehvar et al. (2013)	0.868	0.677	0.457	0.710	0.690	
Zesch et al. (2008)	0.820	—	—	0.710	—	
Collobert and Weston (2008)	0.480	0.610	0.380	—	0.570	
Word2vec (Baroni et al., 2014)	0.840	0.800	0.700	—	0.800	
GloVe	0.769	0.666	0.559	0.577	0.763	
ESA	0.749	—	—	—	—	
PMI-SVD	0.738	0.659	0.523	0.337	0.726	
Word2vec	0.732	0.707	0.476	0.343	0.665	
SENSEMBED <sub>closest</sub>	<b>0.894</b>	0.756	0.645	<b>0.734</b>	0.779	0.770
SENSEMBED <sub>weighted</sub>	0.871	<b>0.812</b>	<b>0.703</b>	0.639	<b>0.805</b>	0.794

# MUFFIN: Multilingual UniFied Flexible Interpretation

[Camacho-Collados et al., ACL 2015]



- Unification is based on the **Wikipedia Bitaxonomy**
- We obtain an **explicit semantic vector** for each BabelNet synset (**multilingual** and **unified!**)
  - Vector components are **concepts** and their values are their importance for the target concept represented by the vector

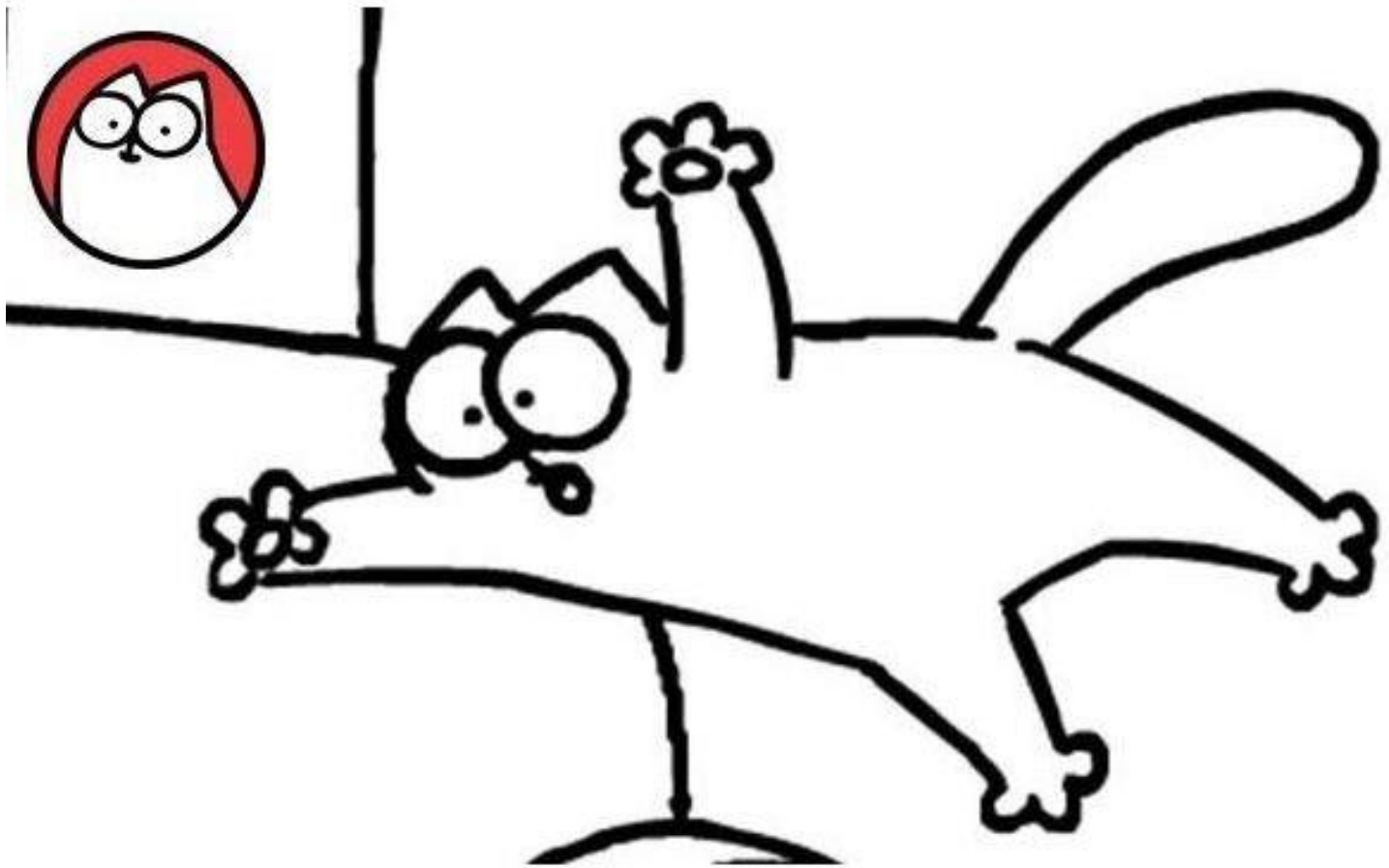
# MUFFIN: Multilingual UniFied Flexible Interpretation [Camacho Collados et al., ACL 2015]

- Performs consistently well across languages:

English	$\rho$	$r$	German	$\rho$	$r$	French	$\rho$	$r$
MUFFIN	0.83	0.84	MUFFIN	<b>0.77</b>	<b>0.76</b>	MUFFIN	<b>0.71</b>	<b>0.77</b>
SOC-PMI	–	0.61	SOC-PMI	–	0.27	SOC-PMI	–	0.19
PMI	–	0.41	PMI	–	0.40	PMI	–	0.34
Retrofitting	0.74	–	Retrofitting	0.60	–	Retrofitting	0.61	–
LSA-Wiki	0.69	0.65	–	–	–	LSA-Wiki	0.52	0.57
Wiki-wup	–	0.59	Wiki-wup	–	0.65			
SSA	0.83	<b>0.86</b>	Resnik	–	0.72			
NASARI	0.84	0.82	Lesk_hyper	–	0.69			
ADW	<b>0.87</b>	0.81						
Word2Vec	–	0.84						
PMI-SVD	–	0.74						
ESA	–	0.72						

Spearman ( $\rho$ ) and Pearson ( $r$ ) correlation performance of different systems on the English, German and French RG-65 datasets.

# CRAZY TIME!!!





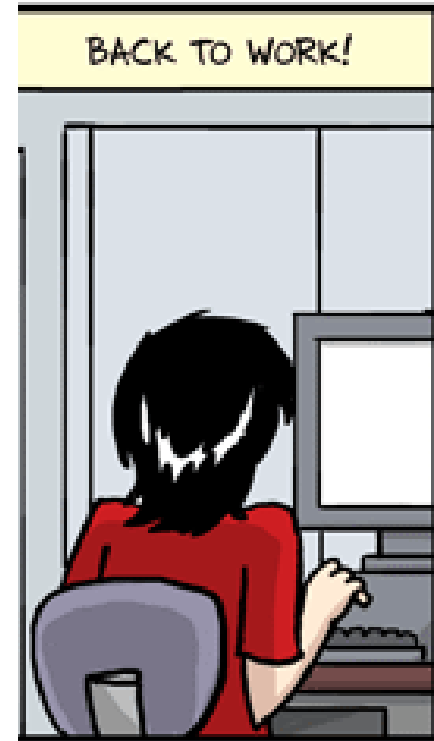
# Now some of you...

will...

..."receive" a BabelNet t-shirt!!!



[model is not included]



WWW.PHDCOMICS.COM

# Summarizing



BabelNet



+ preview on **sense embeddings and explicit multilingual vectors** for state-of-the-art semantic similarity!

**NOW...**

**WHAT WILL YOU DO WITH  
BABELFY TONIGHT?**

Thanks or...





SAPIENZA  
UNIVERSITÀ DI ROMA

**Roberto Navigli**

Linguistic Computing Laboratory

<http://lcl.uniroma1.it>

