Word Sense Disambiguation and Crowdsourcing

Roberto Navigli

Dept. of Computer Science

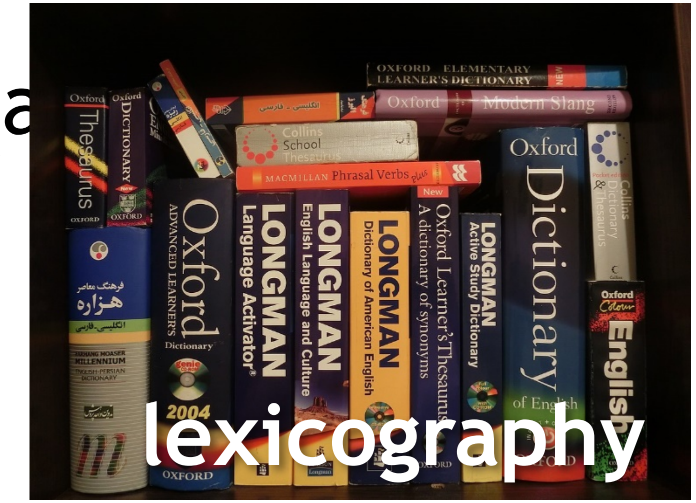
Sapienza University of Rome

February 18th, 2019 - ELEXIS observer event, Vienna

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Lexicographic data for Natural Language Processing and vice versa

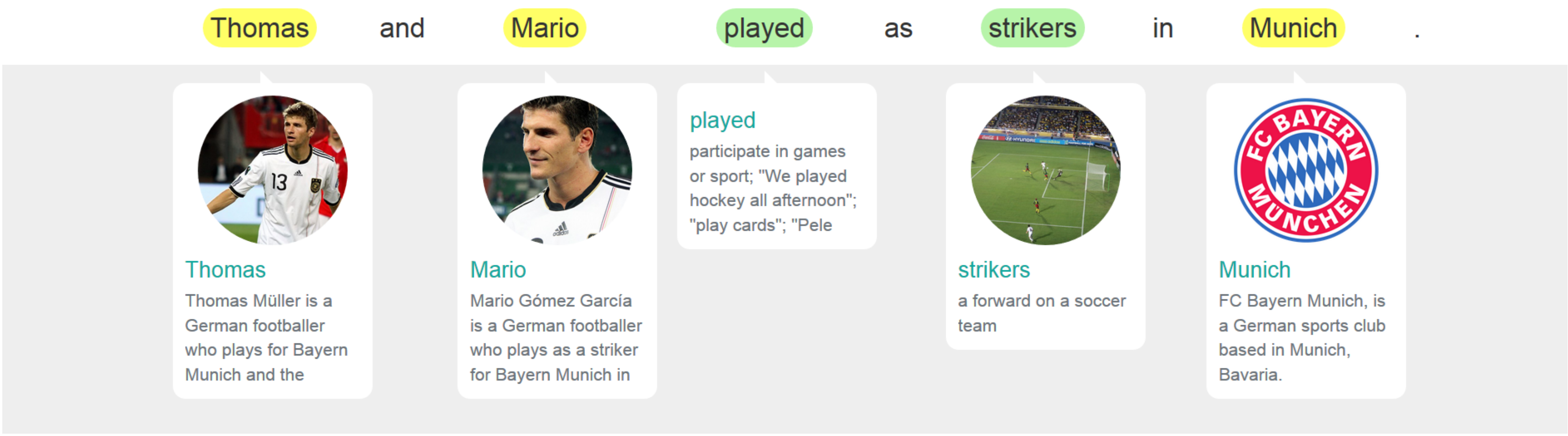
- We aim to show the **impact of lexicographic data for NLP**
- A paradigmatic task is Word Sense Disambiguation thanks to its need to leverage lexical-semantic knowledge resources
- Howev im to show tha **lexicog**



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Multilingual Word Sense Disambiguation and Entity Linking

- **Word Sense Disambiguation:** automatic assignment of senses to words in context
- **Entity Linking:** automatic assignment of named entities to mentions in context



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.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Multilingual Word Sense Disambiguation and Entity Linking

- **Objective 1:** Develop algorithms that will use ELEXIS lexicographic resources to bootstrap disambiguation in a dozen languages
- **Objective 2:** Show high performance in many languages
  - Quantitative evaluation based on standard multilingual datasets (SemEval 2013; 2015 on multilingual WSD; Entity Linking datasets)
  - Perform validation in multiple languages and with different sense inventories: demonstrate **high-quality sense annotations**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Challenges in WSD and Entity Linking

- **Issues:**
  - **The knowledge acquisition bottleneck:**
    - **Supervised approaches** suffer from **lack of annotated data** (only English and little else)
    - **Knowledge-based approaches** need computational lexicons, semantic collocations, graph-like dictionary structure, etc.
  - **Reference inventories**
    - WordNet is too fine grained
    - Wikipedia is too rich
- The ELEXIS dictionary matrix will prove **important benefits for both issues**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

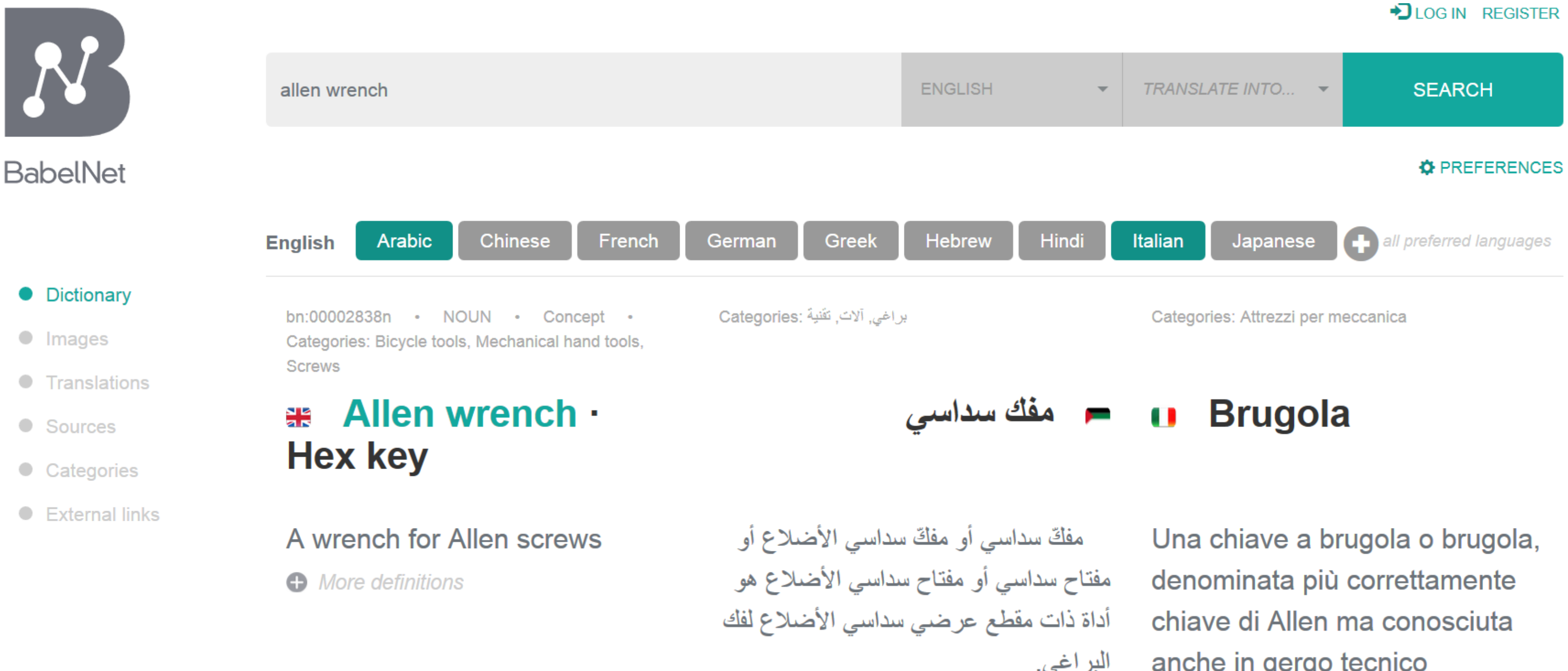
Workplan (1/2)

- **Textual data** from:
  - the Universal Dependencies project (POS tagged)
  - the *TenTen corpora from Lexical Computing
  - Semantically-annotated corpora from partners
- **Phase 1a (October 2018/February 2019):**
  - **Algorithms:** Babelfy (Uniroma1) + Wikifier (JSI)
  - **Inventory:** use existing inventories (BabelNet, Wikipedia)
  - **Validation:** show the data to lexicographers in ELEXIS + observers
  - **Goal:** prepare the framework
- **Phase 1b (February 2019/June 2019):**
  - **Disambiguation** of the corpora + analysis

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.

BabelNet: a shared multilingual inventory of meanings

- **Multilingual**: the same concept in tens of languages
- It **integrates different kinds of open resources**, such as WordNet, Wikipedia, Wikidata, Wiktionary, etc.
- **Wide coverage**: 284 languages and 16 million entries!
- Used by **more than 800 universities and research centers!**



allen wrench

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

Allen wrench · **Hex key**

A wrench for Allen screws

Categories: براغي, آلات, تقنية

مفك سداسي

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

Categories: Attrezzi per meccanica

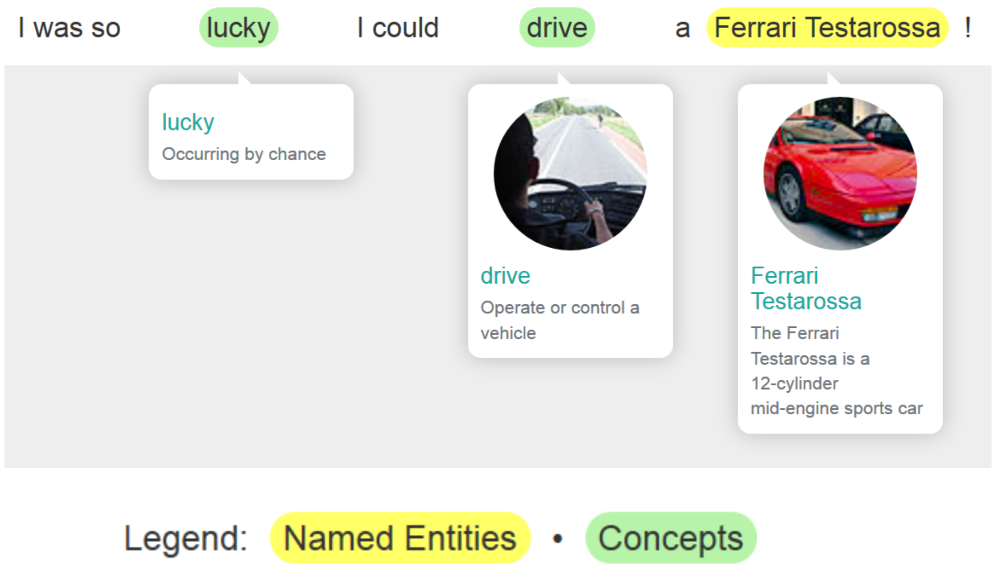
Brugola

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

This project has r

Disambiguation: Babelfy

- We used Babelfy for disambiguating the Wikipedia corpus
- Why?
  - The first (and only) system that performs Word Sense Disambiguation (common nouns, verbs, adjectives, adverbs) and Entity Linking (names) **jointly**



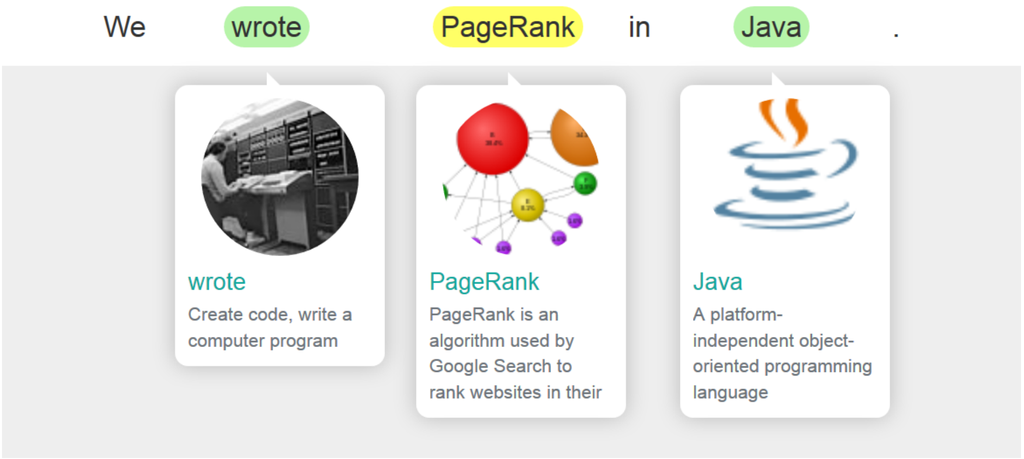
I was so lucky I could drive a Ferrari Testarossa !

lucky
Occurring by chance

drive
Operate or control a vehicle

Ferrari Testarossa
The Ferrari Testarossa is a 12-cylinder mid-engine sports car

Legend: Named Entities • Concepts



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

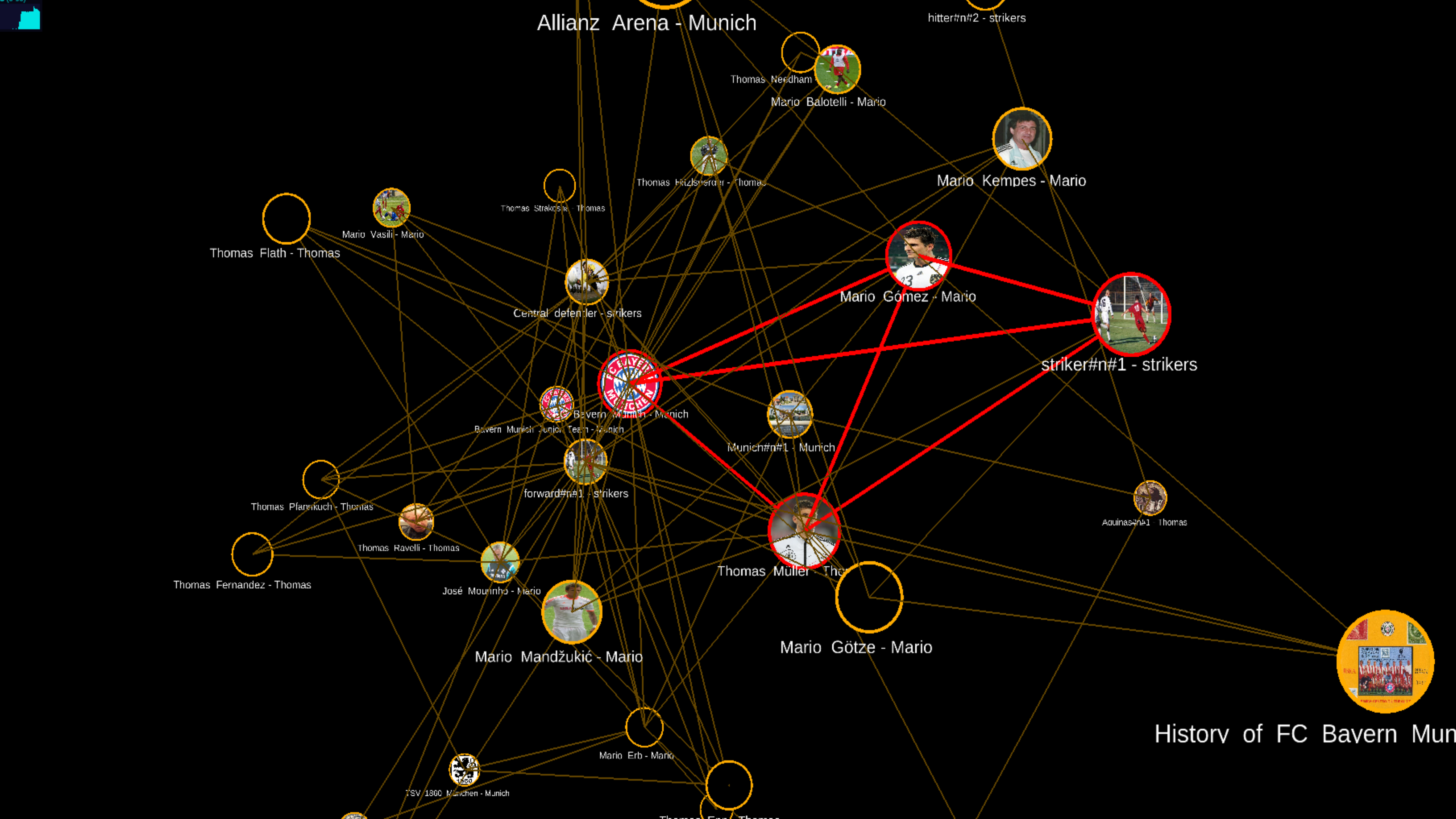
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Disambiguation: Babelfy

- We used Babelfy for disambiguating the Wikipedia corpus
- Why?
  - The first (and only) system that performs Word Sense Disambiguation (common nouns, verbs, adjectives, adverbs) and Entity Linking (names) **jointly**
  - **Knowledge-based:** does not need millions of sentences annotated in each language
  - Works in **arbitrary languages** (284 languages)
  - Can disambiguate **texts written in mixed languages** (language-agnostic setting)

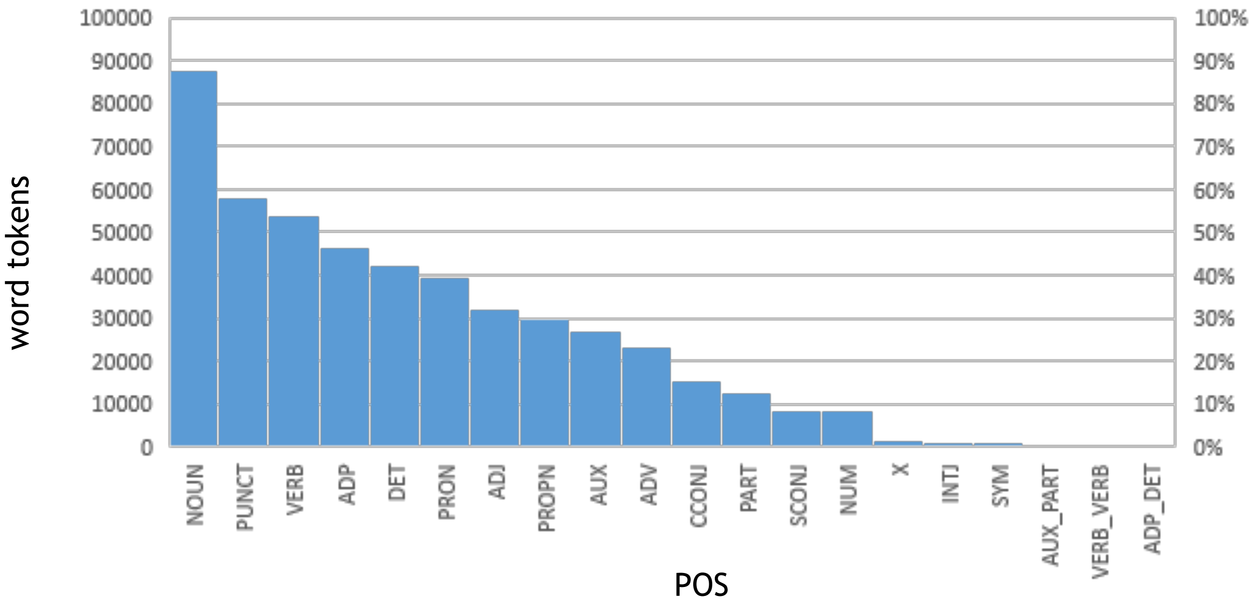
"Word sense disambiguation and entity linking together!"

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.


Allianz Arena - Munich
hitter#n#2 - strikers
Thomas Needham
Mario Balotelli - Mario
Mario Kempes - Mario
Thomas Hitzlsperger - Thomas
Thomas Strakosha - Thomas
Mario Vasili - Mario
Thomas Flath - Thomas
Central defender - strikers
Mario Gómez - Mario
striker#n#1 - strikers
FC Bayern Munich - Munich
Bayern Munich Junior Team - Munich
Munich#n#1 - Munich
forward#n#1 - strikers
Thomas Pfannkuch - Thomas
Thomas Ravelli - Thomas
Aquinas#n#1 - Thomas
Thomas Müller - Tho
Thomas Fernandez - Thomas
José Mourinho - Mario
Mario Mandžukić - Mario
Mario Götze - Mario
History of FC Bayern Mun
Mario Erb - Mario
TSV 1860 München - Munich
Thomas Erb - Thomas

Statistics on the Babelfied English Datasets from UD

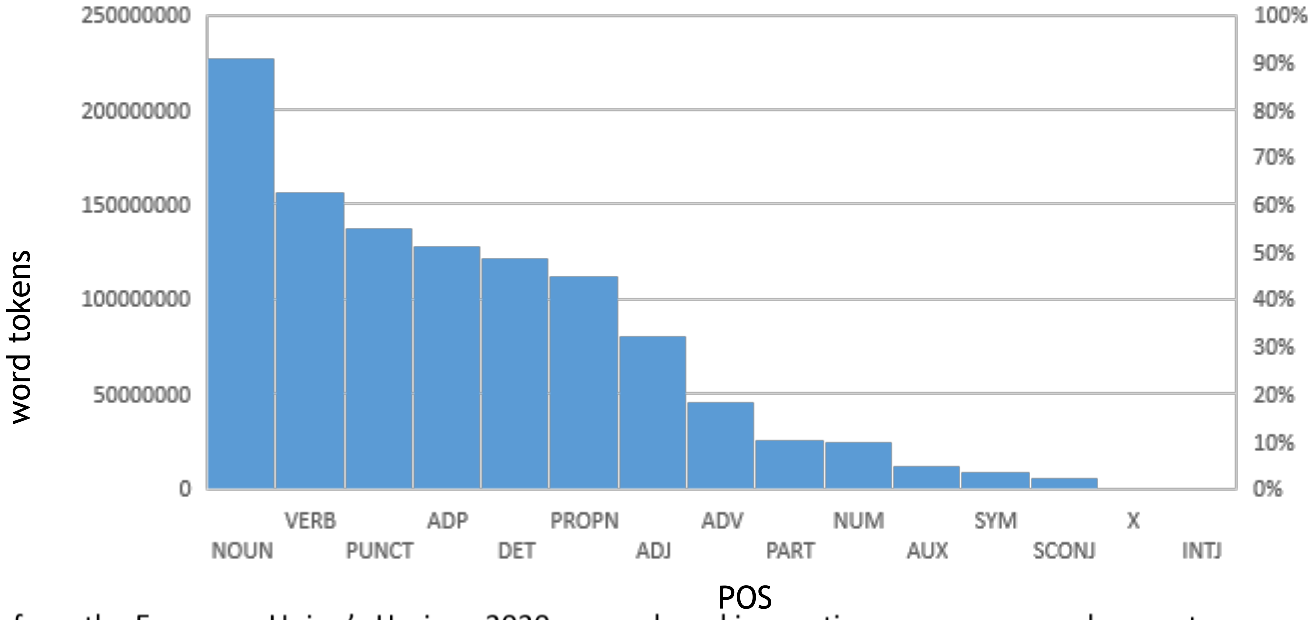
- Total number of word tokens: 488515
- Total number of word types: 26892
- Total number of disambiguated word tokens: 85094
- Total number of disambiguated word types: 24144



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

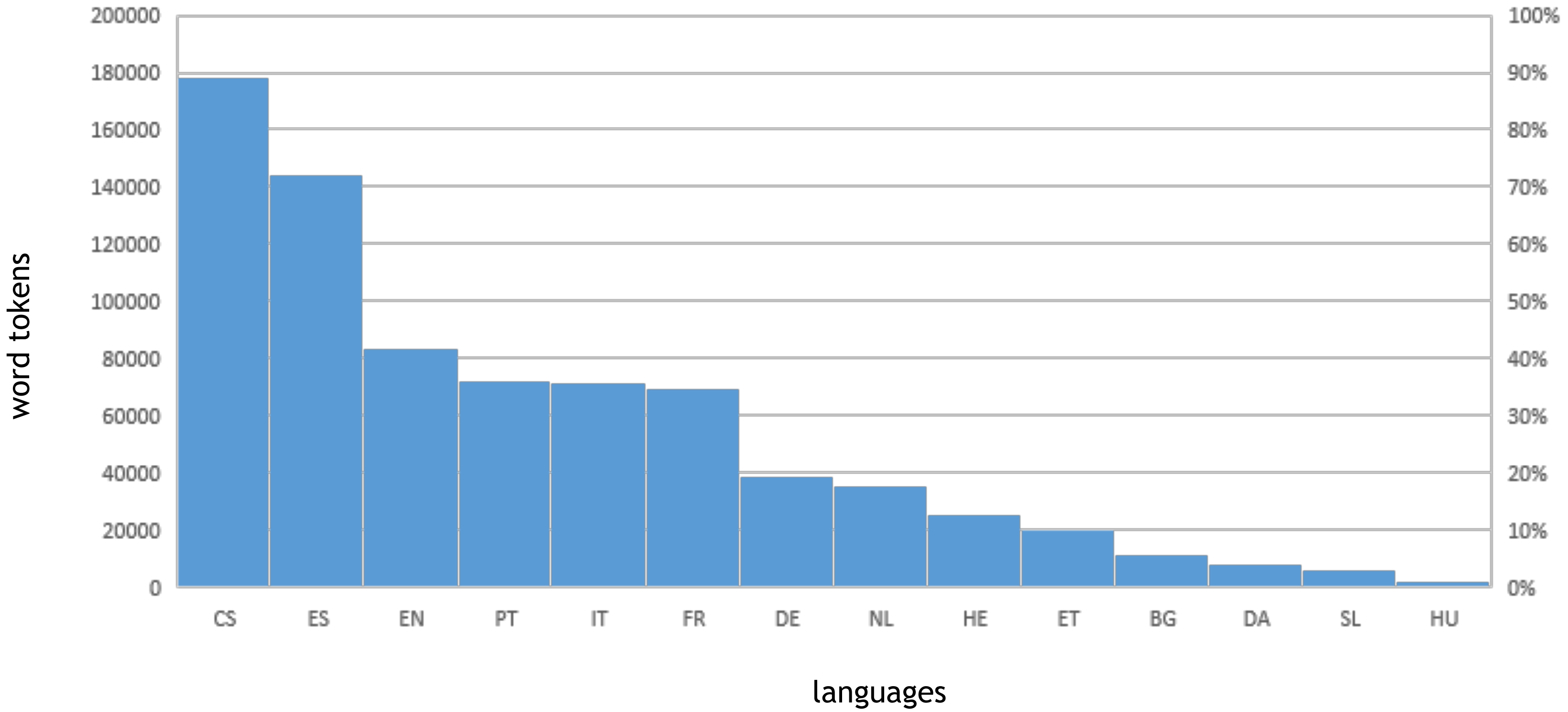
Statistics on the Babelfied English Datasets from the *TenTen Corpora

- Total number of word tokens: 1170202338
- Total number of word types: 5454205
- Total number of disambiguated word tokens: 295783220
- Total number of disambiguated word types: 844705



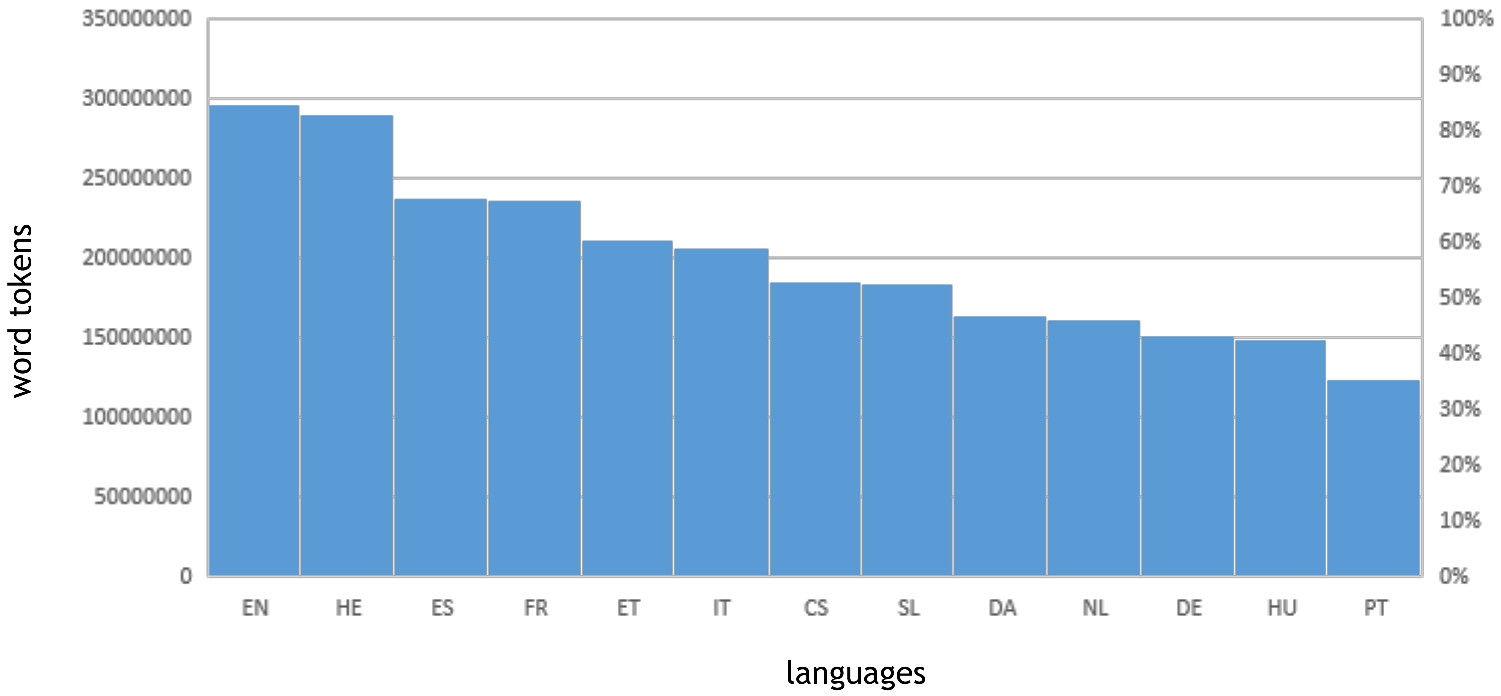
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Overall statistics on words Babelfied across languages of UD Corpora



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Overall statistics on words Babelfied across languages of *TenTen Corpora



This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.

Qualitative analysis + decisions about lexicographic evaluation

- We are in the process of defining the evaluation process
- Shall we check if the best option was chosen among those in the inventory?
- Annotations (potential options: correct, so and so, wrong?)
- Evaluation measures

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

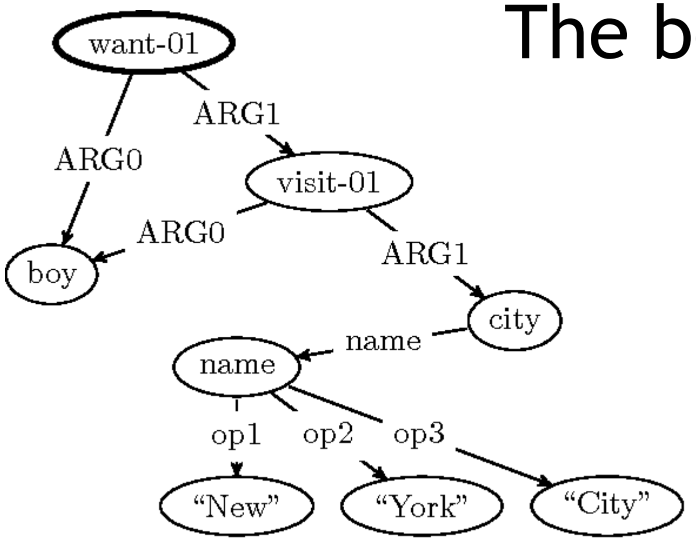
Workplan (2/2)

- **Phase 2 (June 2019-2021):**
  - **Algorithms:** New algorithm with multilingual sense embeddings + semantic graphs (Uniroma1)
  - **Inventory:** use the ELEXIS dictionary matrix from WP2
  - **Validation:** show the data to lexicographers in ELEXIS + observers
  - **Goals:**
    1. show we can now disambiguate in arbitrary languages with reputable dictionaries
    2. show improvements coming from the dictionary matrix resulting from WP2

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Multilingual semantic parsing

- Semantic parsing is the task of mapping sentences to a formal representation
  - Abstract Meaning Representation (AMR)
  - Universal Conceptual Cognitive Annotation (UCCA)
  - CCG-based like Discourse Representation Structures (DRS)



The boy wants to visit New York City

```
(w / want-01
  :ARG0 (b / boy)
  :ARG1 (g / visit-01
           :ARG0 b
           :ARG1 (c / city
                    :name (n / name
                             :op1 "New"
                             :op2 "York"
                             :op3 "City"))))
```

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

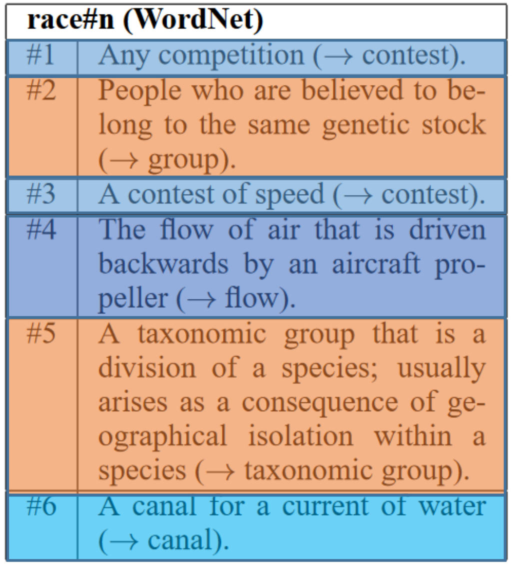
Multilingual semantic parsing

- Semantic parsing is the task of mapping sentences to a formal representation
- **Objective 1:** develop algorithms for semantic parsing in multiple languages which take advantage of ELEXIS lexicographic data
- **Objective 2:** exploit bilingual and multilingual data to innovate semantic parsing algorithms
- **Expectations from other partners:** creation of a multilingual test set benchmark (à la SemEval) for the task; curation/ validation of verb frames for parsing in different languages based on ELEXIS data

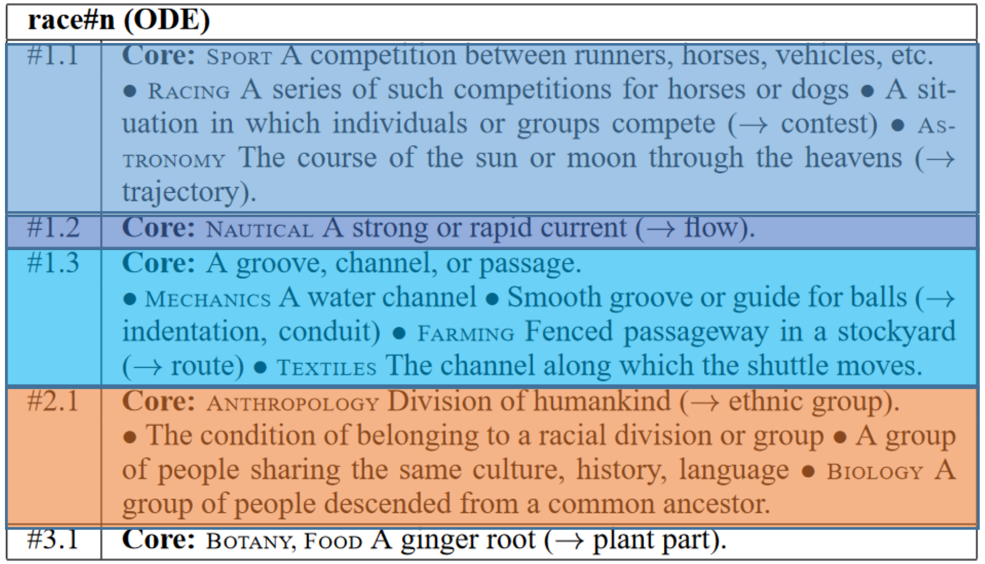
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Lexical-semantic analytics for NLP

- Based on analytics computed on the ELEXIS resources for words, phrases, collocations, senses, domains, etc. we will explore three directions:
  - **T3.3.1 Sense clustering:** semi-automatic algorithms to group fine-grained sense distinctions, also across languages



race#n (WordNet)	
#1	Any competition (→ contest).
#2	People who are believed to belong to the same genetic stock (→ group).
#3	A contest of speed (→ contest).
#4	The flow of air that is driven backwards by an aircraft propeller (→ flow).
#5	A taxonomic group that is a division of a species; usually arises as a consequence of geographical isolation within a species (→ taxonomic group).
#6	A canal for a current of water (→ canal).

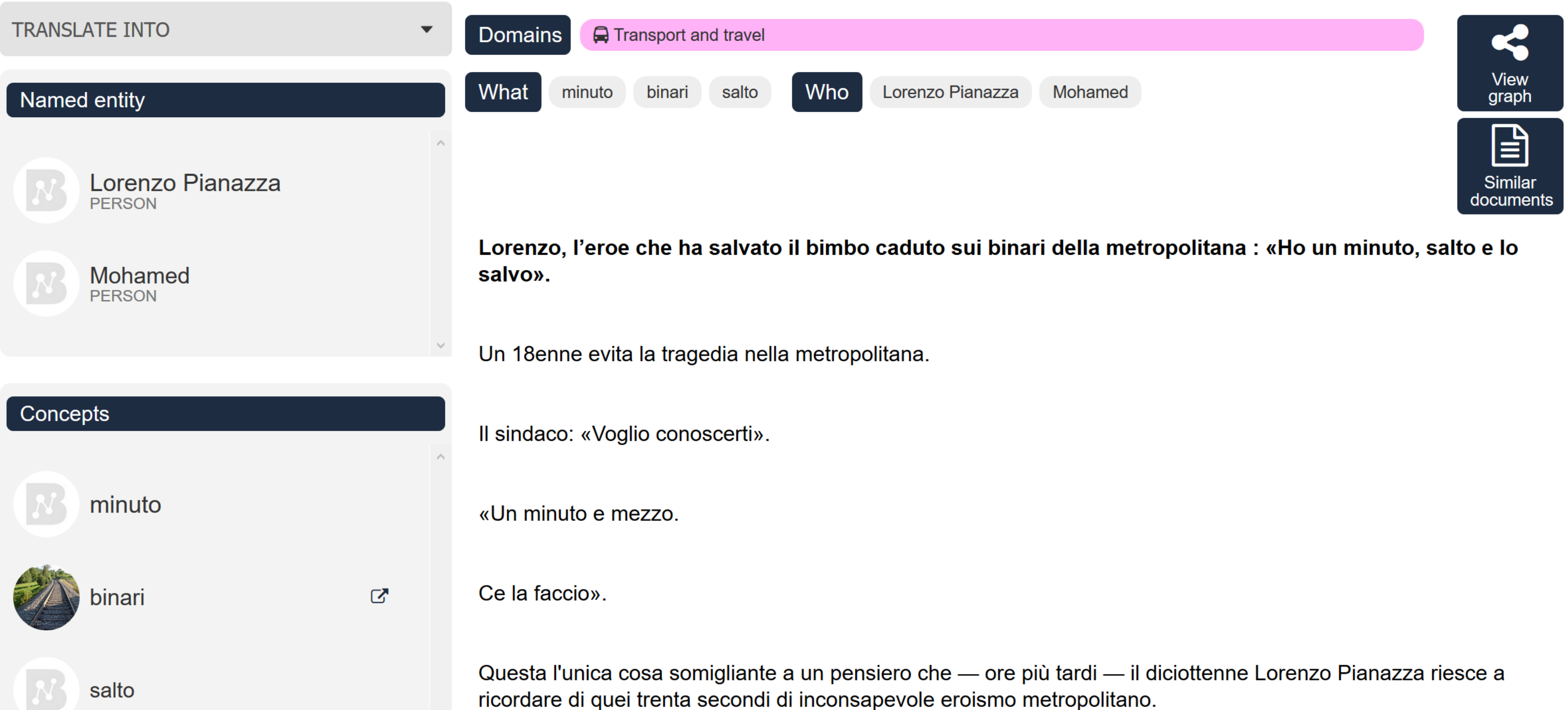


race#n (ODE)	
#1.1	**Core:** Sport A competition between runners, horses, vehicles, etc. • Racing A series of such competitions for horses or dogs • A situation in which individuals or groups compete (→ contest) • Astronomy The course of the sun or moon through the heavens (→ trajectory).
#1.2	**Core:** Nautical A strong or rapid current (→ flow).
#1.3	**Core:** A groove, channel, or passage. • Mechanics A water channel • Smooth groove or guide for balls (→ indentation, conduit) • Farming Fenced passageway in a stockyard (→ route) • Textiles The channel along which the shuttle moves.
#2.1	**Core:** Anthropology Division of humankind (→ ethnic group). • The condition of belonging to a racial division or group • A group of people sharing the same culture, history, language • Biology A group of people descended from a common ancestor.
#3.1	**Core:** Botany, Food A ginger root (→ plant part).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Lexical-semantic analytics for NLP

- Based on analytics computed on the ELEXIS resources for words, phrases, collocations, senses, domains, etc. we will explore three directions:
  - **T3.3.2 Domain labeling of text:** ELEXIS resources shown to improve domain labeling across languages



Lorenzo, l’eroe che ha salvato il bimbo caduto sui binari della metropolitana : «Ho un minuto, salto e lo salvo».

Un 18enne evita la tragedia nella metropolitana.

Il sindaco: «Voglio conoscerti».

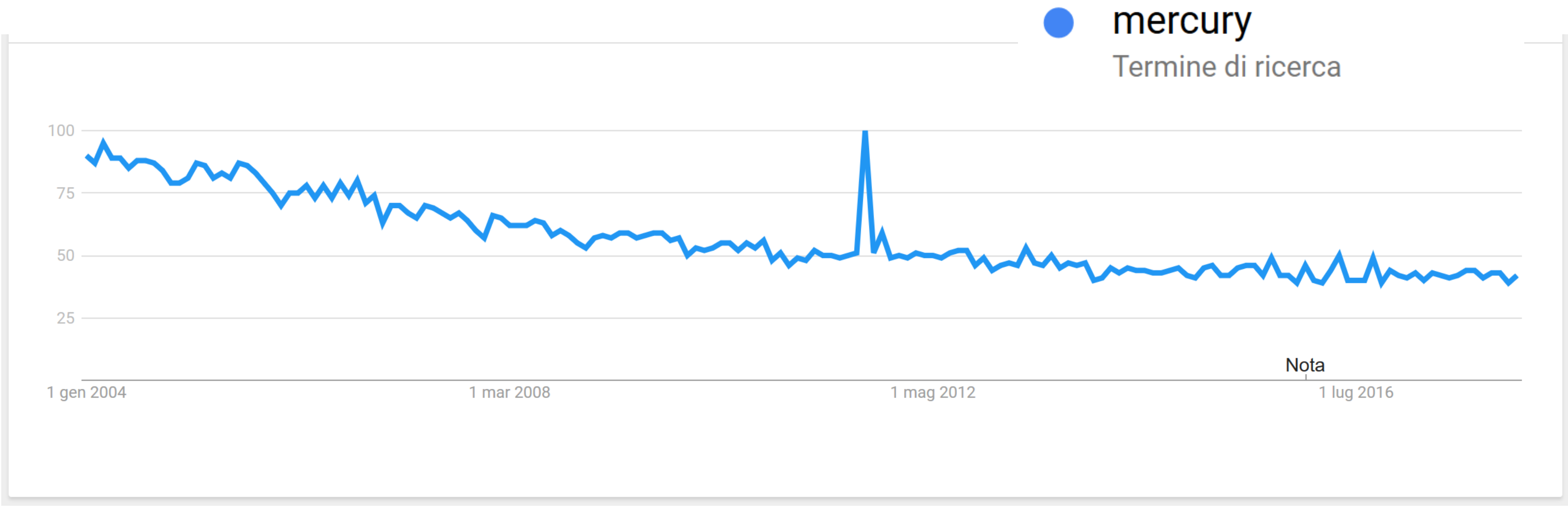
«Un minuto e mezzo.

Ce la faccio».

Questa l'unica cosa somigliante a un pensiero che — ore più tardi — il diciottenne Lorenzo Pianazza riesce a ricordare di quei trenta secondi di inconsapevole eroismo metropolitano.

Lexical-semantic analytics for NLP

- Based on analytics computed on the ELEXIS resources for words, phrases, collocations, senses, domains, etc. we will explore three directions:
  - **T.3.3.3 Diachronic distribution of senses:** sense frequency ranking over time across resources (Most Frequent Sense is a strong baseline in WSD)



This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.

Challenges in lexical-semantic analytics

- **Sense clustering:**
  - Fine granularity
  - Not obvious what a good cluster of senses is
- **Domain labeling of text:**
  - Elicit information from ELEXIS resources (what is a good set of domain labels? which resources provide domain-specific content?)
  - Work in dozens of languages
- **Diachronic distribution of senses:**
  - Create reliable distributions of senses in many languages
  - Leverage such distributions in WSD and Entity Linking
- **See interaction with WP4**

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

Crowdsourcing and gamification

- Objectives:
  - Validating the output of WP2 (links between resources, the dictionary matrix)
  - Validating and improving the data produced by WP3 and WP4

- Proposal for a crossword game developed jointly with Babelscape
- Other crowdsourcing efforts are on-going

- Goal: collect experiences from the consortium and observers

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 731015.

Questions?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.