A feasibility study of a Spanish version of the UMLS

Àlex Bravo, Pablo Accuosto & Horacio Saggion

red.es

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y EMPRESA

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

upf. Universitat Pompeu Fabra Barcelona

taln

Overview

1. Overview of the Unified Medical Language System (UMLS)
2. Spanish UMLS Vs English UMLS
3. Biomedical Resources (Corpora and Tools) and Processing and Analysing Corpora
4. Methods to expand the Spanish terminology
5. Results

upf. taln

Grup de Recerca en
Tractament Automàtic
del Llenguatge Natural

Automatic Natural
Language Processing
Research Group

What is the UMLS?

What is the UMLS?

Developed by the National Library of Medicine (USA), the UMLS is a system which facilitates the development of computer systems in the health and biomedicine.

It connects several terminologies in the **health and biomedical vocabularies and standards** to enable interoperability between computer systems.

The UMLS integrates 154 terminological resources for 25 languages:

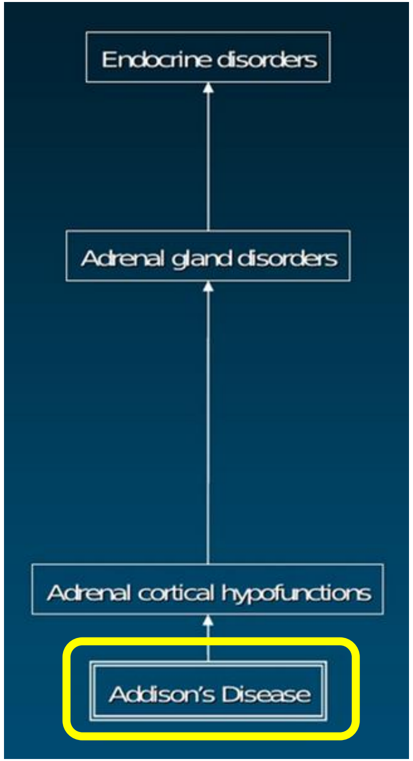
- 133 in English
- 9 in Spanish
- and 1 in Basque

Some of the terminological sources in English are:

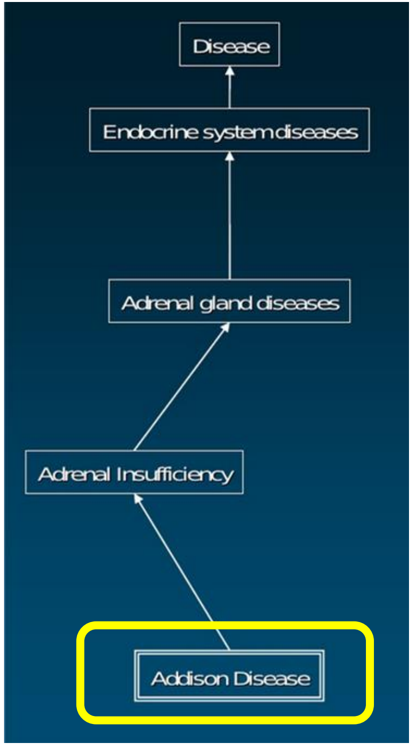
- MeSH: Medical Subject Headings (scientific articles / books)
- SNOMED CT: Clinical Healthcare Terminology
- MedDRA: Regulatory information and clinical safety data for human medical products

What is the UMLS?

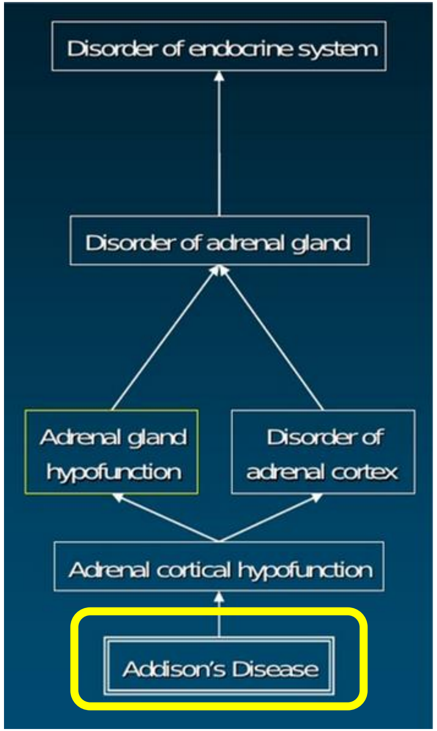
MEDDRA



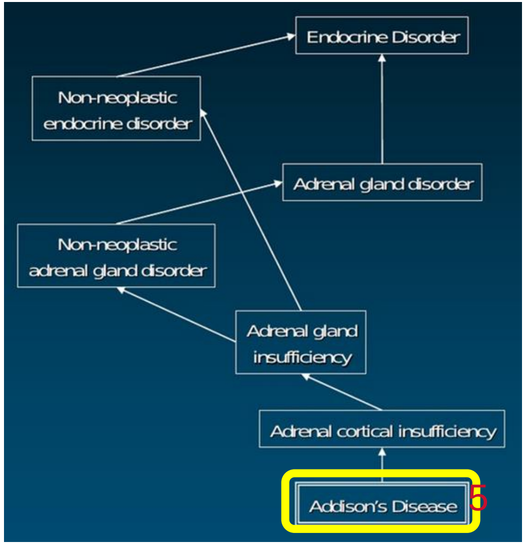
MESH



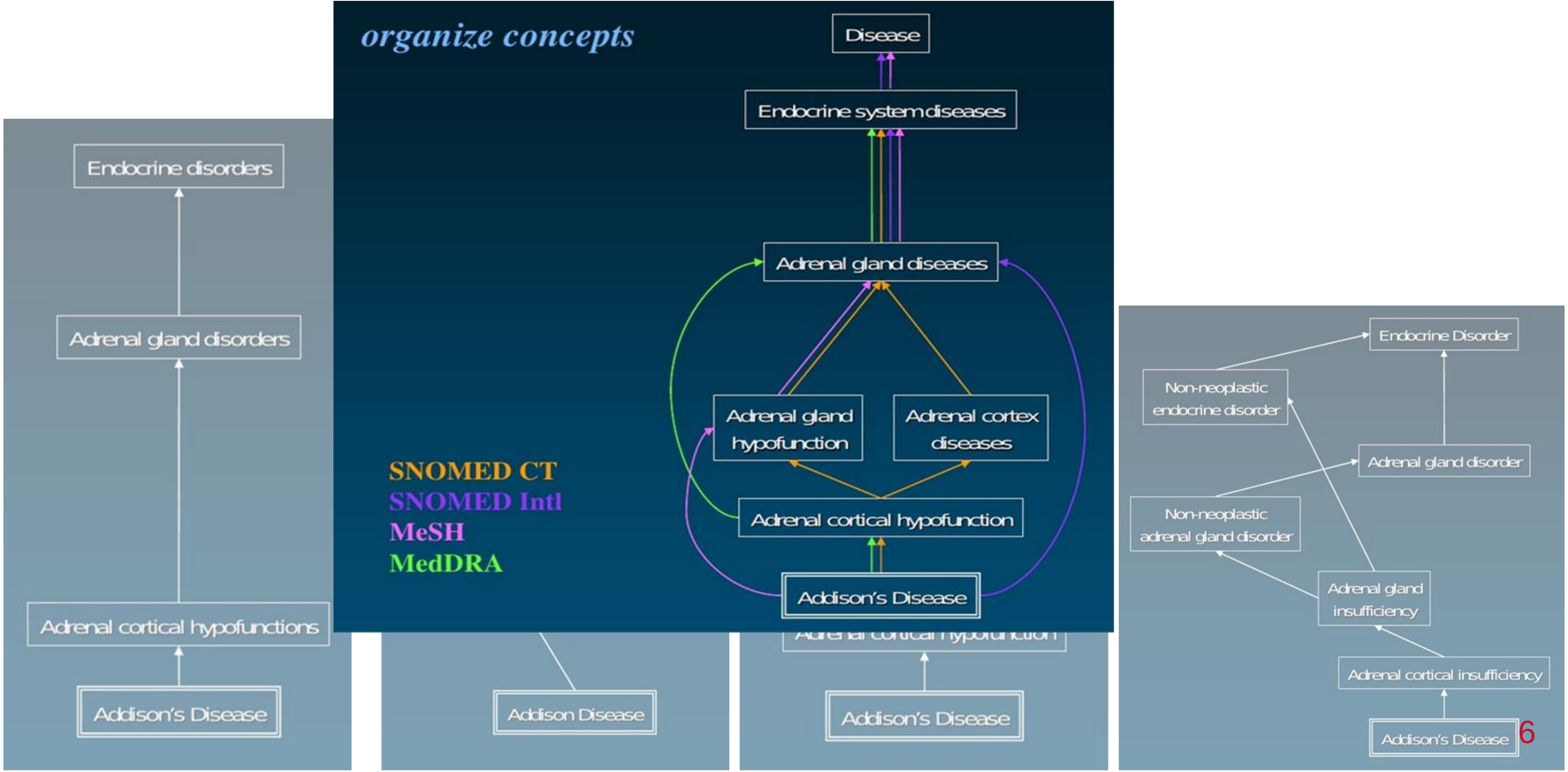
SNOMED CT


Disorder of endocrine system
Disorder of adrenal gland
Adrenal gland hypofunction
Disorder of adrenal cortex
Adrenal cortical hypofunction
Addison's Disease

NCI THESAURUS


Endocrine Disorder
Non-neoplastic endocrine disorder
Adrenal gland disorder
Non-neoplastic adrenal gland disorder
Adrenal gland insufficiency
Adrenal cortical insufficiency
Addison's Disease

What is the UMLS?


organize concepts
Disease
Endocrine system diseases
Adrenal gland diseases
Adrenal gland hypofunction
Adrenal cortex diseases
Adrenal cortical hypofunction
Addison's Disease
SNOMED CT
SNOMED Intl
MeSH
MedDRA
Endocrine disorders
Adrenal gland disorders
Adrenal cortical hypofunctions
Addison's Disease
Addison Disease
Adrenal cortical hypofunction
Addison's Disease
Endocrine Disorder
Non-neoplastic endocrine disorder
Adrenal gland disorder
Non-neoplastic adrenal gland disorder
Adrenal gland insufficiency
Adrenal cortical insufficiency
Addison's Disease

What is the UMLS?

The UMLS is composed of:

- Metathesaurus:
    - The largest thesaurus in the biomedical domain
    - Terminology from different biomedical resources
    - It assigns a Concept Unique Identifier (CUI) to the terms that denote the same concept
        - **C0020538** → ‘High blood pressure’, ‘Systemic arterial hypertension’ and ‘Hypertensive vascular disease’.

- Semantic Network:
    - Organizes the concepts with categories (**Semantic Types**)
    - And relations between them

- SPECIALIST Lexicon:
    - Composed of lexical items including POS and variant information **(only in English)**

upf. taln Grup de Recerca en Tractament Automàtic del Llenguatge Natural Automatic Natural Language Processing Research Group

Spanish UMLS Vs English UMLS

Spanish UMLS Vs English UMLS

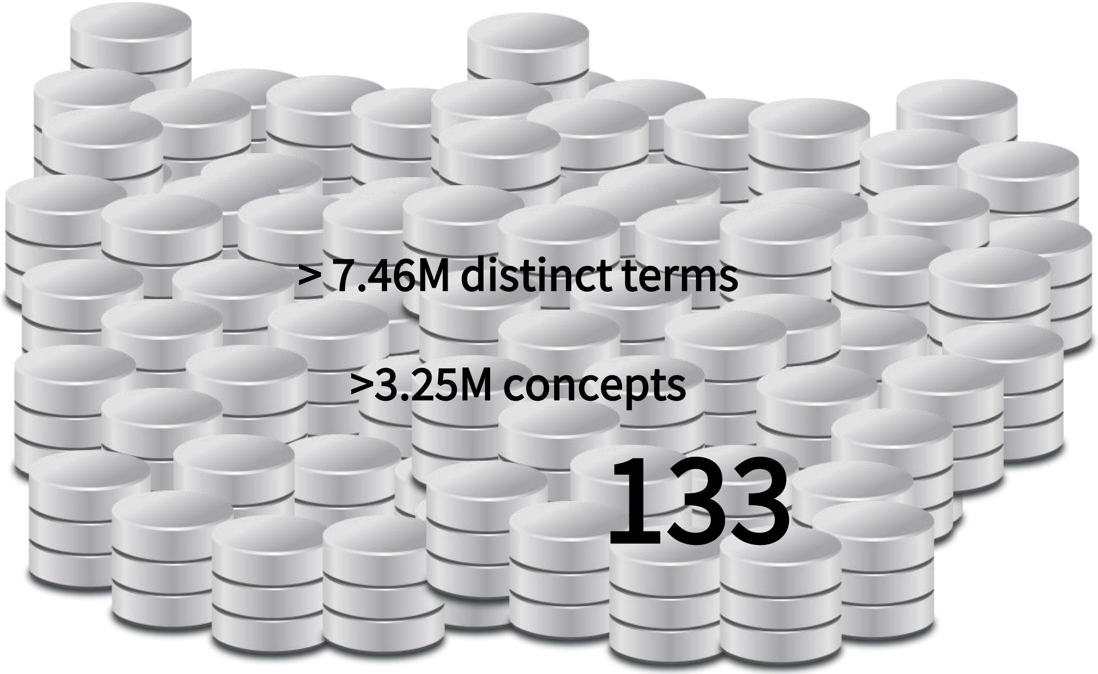
The Spanish UMLS is composed of 9 resources:

> 1.25M distinct terms

>450K concepts

9

UMLS SPA



> 7.46M distinct terms

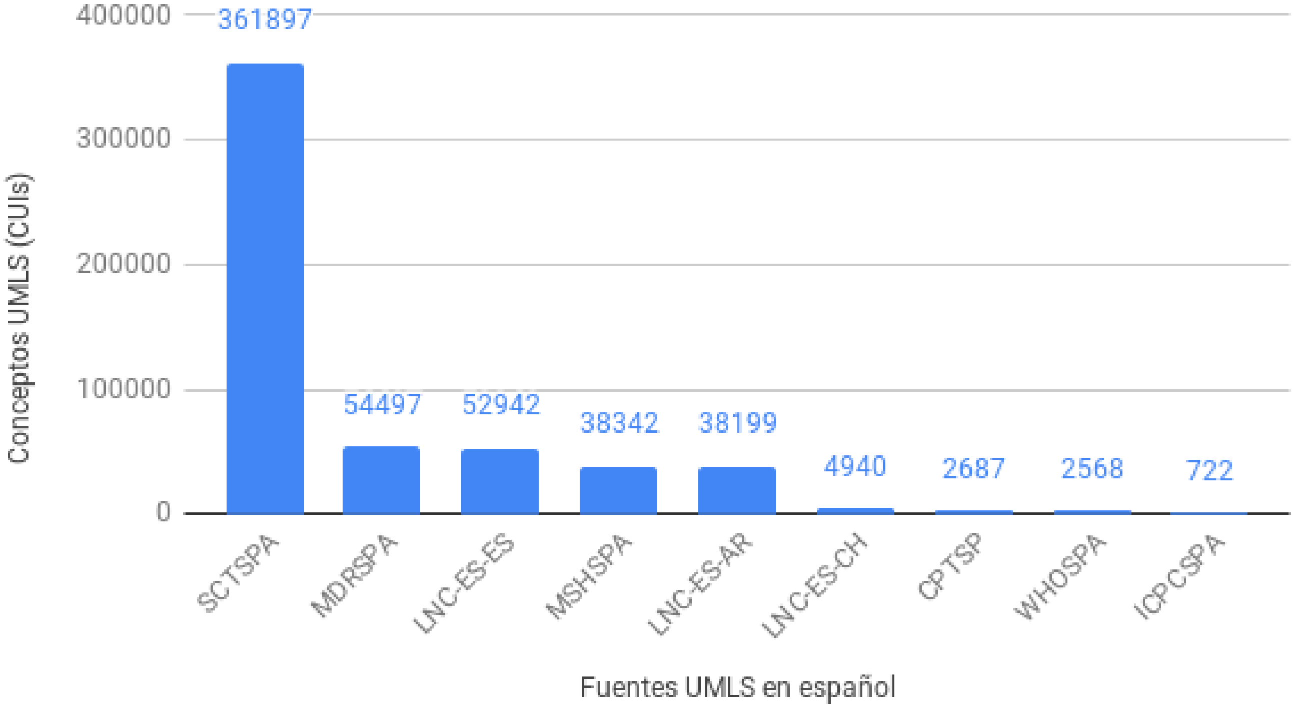
>3.25M concepts

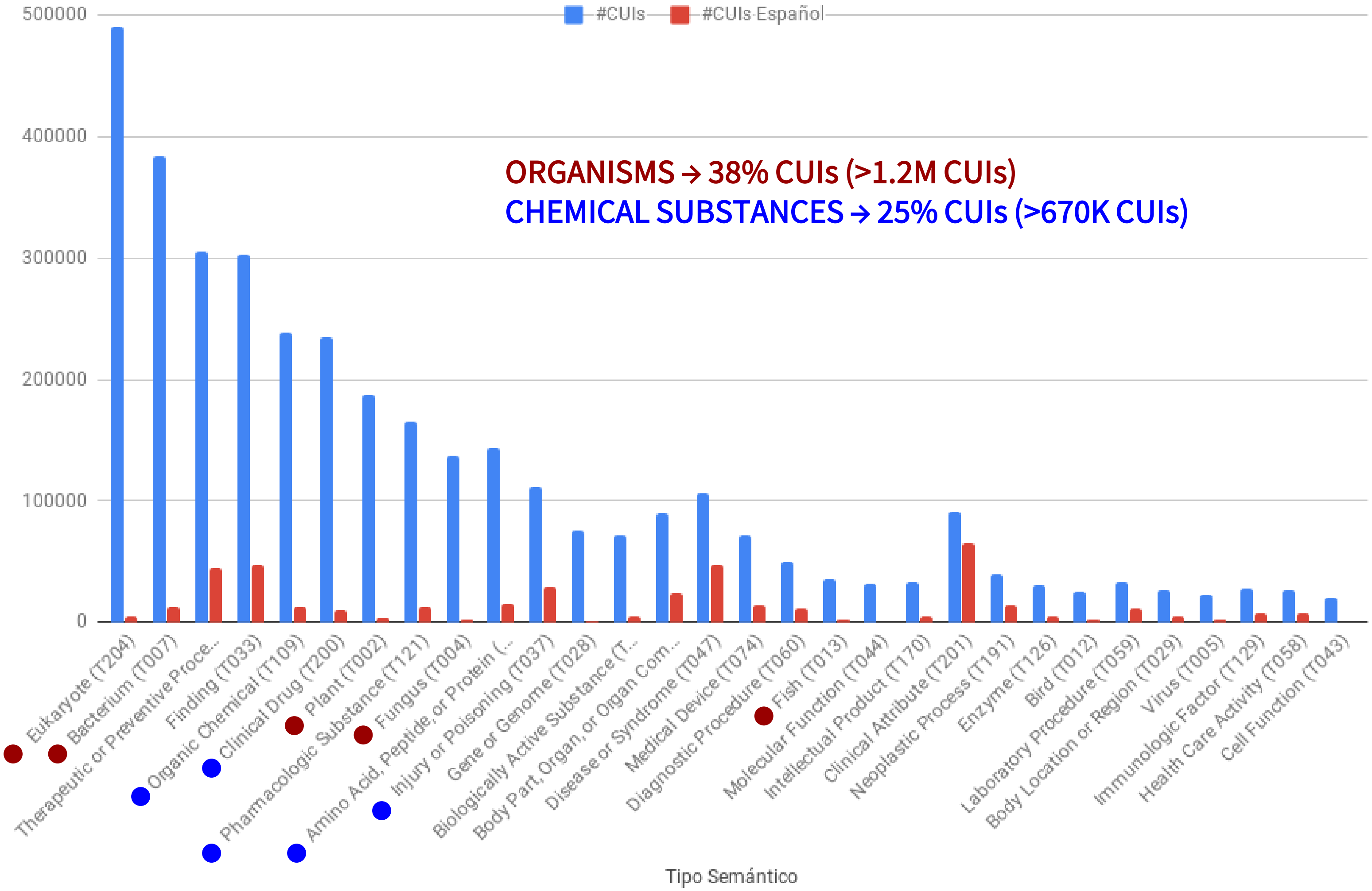
133

UMLS ENG

Spanish UMLS Vs English UMLS

9 Spanish Resources: SCTSPA, MDRSPA, LCN-ES-ES, LNC-ES-AR,LNCS-ES-CH, WHOSPA….





ORGANISMS → 38% CUIs (>1.2M CUIs)

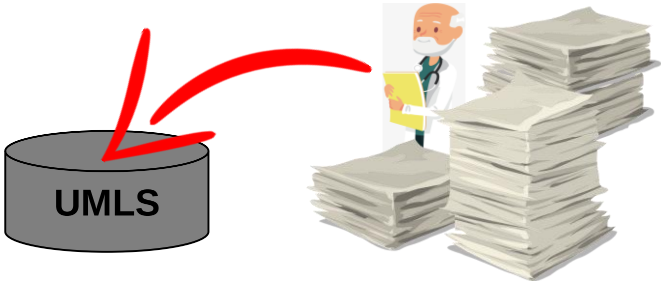
CHEMICAL SUBSTANCES → 25% CUIs (>670K CUIs)

Spanish UMLS Vs English UMLS

We can apply MT to translate the English UMLS.

The UMLS integrates terminology from **curated** biological databases.

- This terminology is **extracted** from biomedical text.
- The terminology in the UMLS also define **how concepts are mentioned** by the authors.



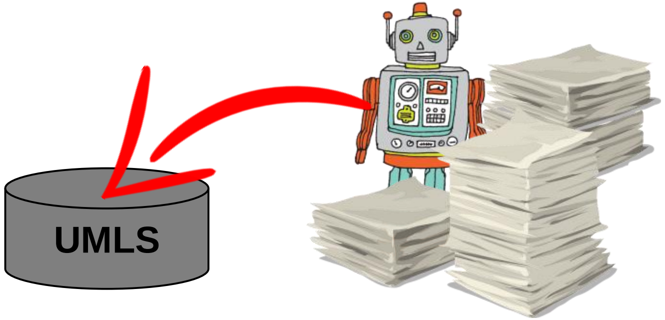
Scientific Publications
Laboratory Reports
Patents
Clinical Reports
Experts

Spanish UMLS Vs English UMLS

We can apply MT to translate the English UMLS.

The UMLS integrates terminology from **curated** biological databases.

- This terminology is **extracted** from biomedical text.
- The terminology in the UMLS also define **how concepts are mentioned** by the authors.
- **GOAL: To extract terminology from biomedical text in Spanish**



Scientific Publications
Laboratory Reports
Patents
Clinical Reports
~~Experts~~ Automatically

upf. taln

Grup de Recerca en
Tractament Automàtic
del Llenguatge Natural

Automatic Natural
Language Processing
Research Group

Resources and Tools

Biomedical Resources

Explore Biomedical Resources → Spanish

- > 2,450 Spanish journals
- Repositories of Journals: IBECS (Índice Bibliográfico Español de Ciencias de la Salud), MEDES (MEDicina en ESpañol), IME (Índice Médico Español), CUIDEN Database, ...
- Search Engines : SciELO, Redalyc, Dialnet, Redib, ...
- Multilingual Corpora: Mantra Gold Standard Corpus, IULA, MedlinePlus, ...
- NLP Tools: FreelingMed, IXA Pipes, META Map, Spanish META Map, ...

Biomedical Resources

Experiments with tools for biomedical entity extraction based on UMLS

- **MetaMap** for English text.
- **UMLS Mapper** for Spanish text (http://www.vicomtech.org/).

Processed Parallel Corpora:

- COPPA → Biomedical Patents
- **MedlinePlus** → Medical Articles
- Scielo → Abstracts from Scientific Publications

Analysis of extracted concepts (CUIs) from English and Spanish

- Statistics & Comparative analysis
- Insights

upf. taln

Grup de Recerca en
Tractament Automàtic
del Llenguatge Natural

Automatic Natural
Language Processing
Research Group

Methods to expand the Spanish terminology

Methods to expand the Spanish terminology

After the analysis of the different resources and datasets:

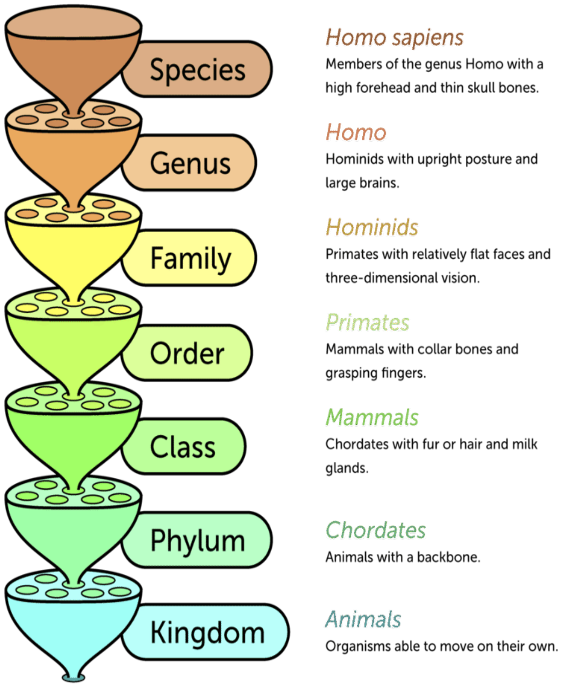
- English UMLS >>> Spanish UMLS
  - ORGANISMS > 40% CUIs
  - CHEMICAL SUBSTANCES > 25% CUIs
- We can process biomedical text in English and Spanish
  - Terminology extraction

Experiment with automatic techniques to expand Spanish terminology.

Methods to expand the Spanish terminology

Transfer via morphology using Knowledge Bases

- Scientific Nomenclature
  - Most organism are associated with a scientific name in Latin according to their Taxonomy
- UMLS → NCBI Taxonomy
  - SCN (Scientific Name) until the **Species** Group (‘Canis lupus’)
  - + 1.2M Concepts
  - **Wikispecies** is indexed by SCN
    - Contains SCN & Common Names in multiple languages: English (wolf), Spanish (lobo), Catalan (llob), Galician (lobo), Asturian (llobu) and Basque (otso)
  - **Multilingual Central Repository** is indexed by Common Names
    - Contains names and synonyms in multiple languages:
      - Spanish, Catalan and Basque
  - **BabelNET** multilingual repository
  - **WordReference** → Synonyms & Inflections (in Spanish: loba, lobos, lobas...)

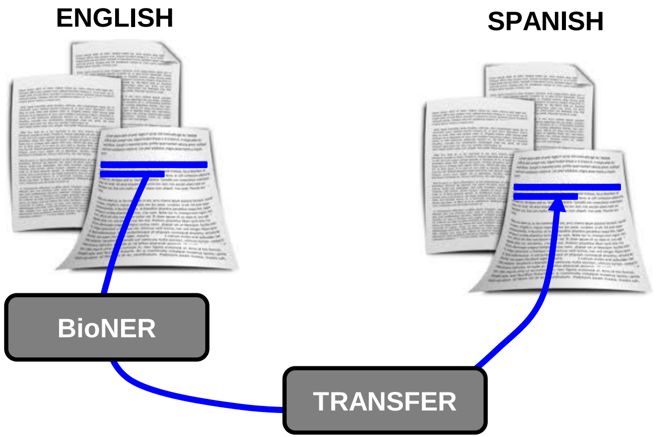

Species
Genus
Family
Order
Class
Phylum
Kingdom
Homo sapiens
Members of the genus Homo with a high forehead and thin skull bones.
Homo
Hominids with upright posture and large brains.
Hominids
Primates with relatively flat faces and three-dimensional vision.
Primates
Mammals with collar bones and grasping fingers.
Mammals
Chordates with fur or hair and milk glands.
Chordates
Animals with a backbone.
Animals
Organisms able to move on their own.

Methods to expand the Spanish terminology

Transfer via language models from biomedical text

Face two limitations:

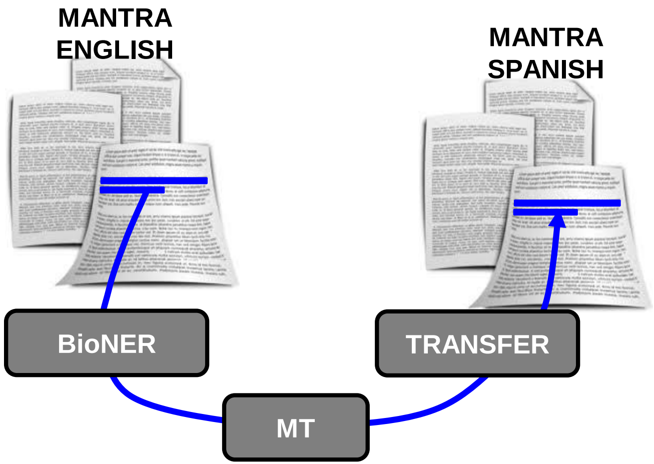
- **Low number** of biomedical resources (vocabularies, tools…) in Spanish.
- The extraction of **novel** biomedical terminology from Spanish text.



Methods to expand the Spanish terminology

Transfer via language models from biomedical text

- Parallel Corpus English-Spanish → Mantra
  - Two datasets: EMEA and MEDLINE
  - With biomedical entity annotations → UMLS
  - Extraction of ~~novel~~ Spanish terminology.
    - Simulation & Evaluation



- Biomedical term extractor → MetaMap
  - Linking terms to UMLS CUI (English)

- Machine Translation → DeepL

- Terminology Transfer → Word embeddings (**FASTTEXT** (https://fasttext.cc/))

Methods to expand the Spanish terminology

Methods to expand the Spanish terminology

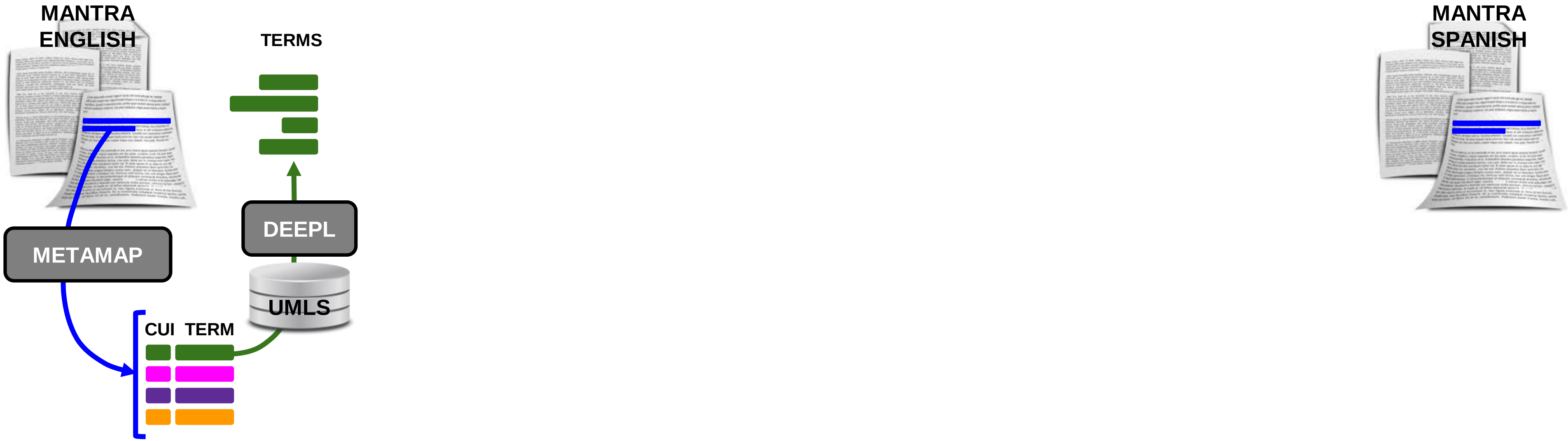
MANTRA ENGLISH

METAMAP

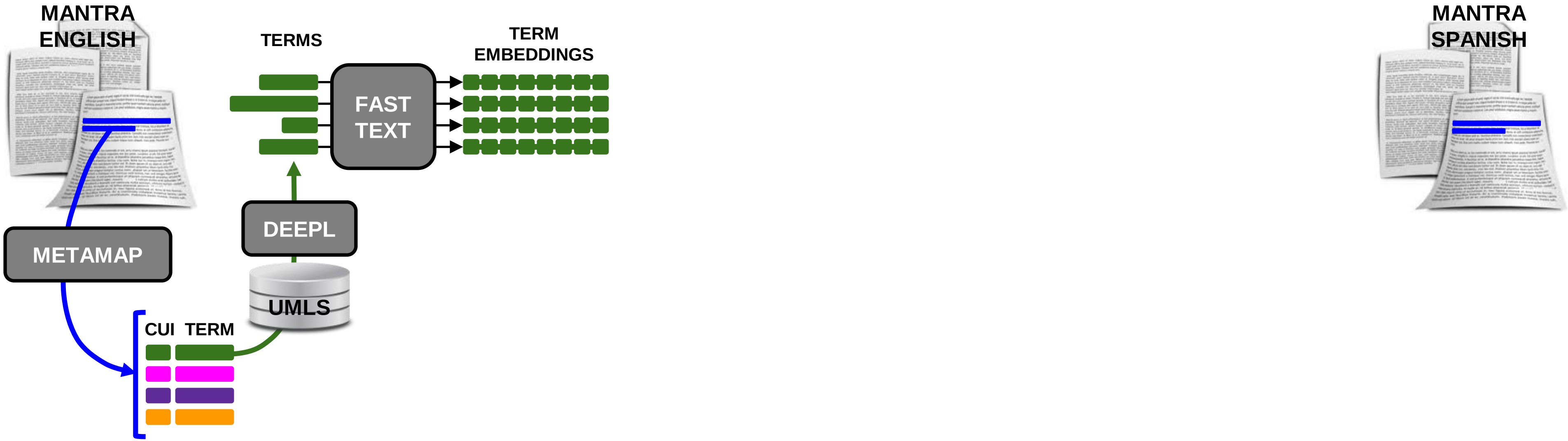
CUI TERM

MANTRA SPANISH

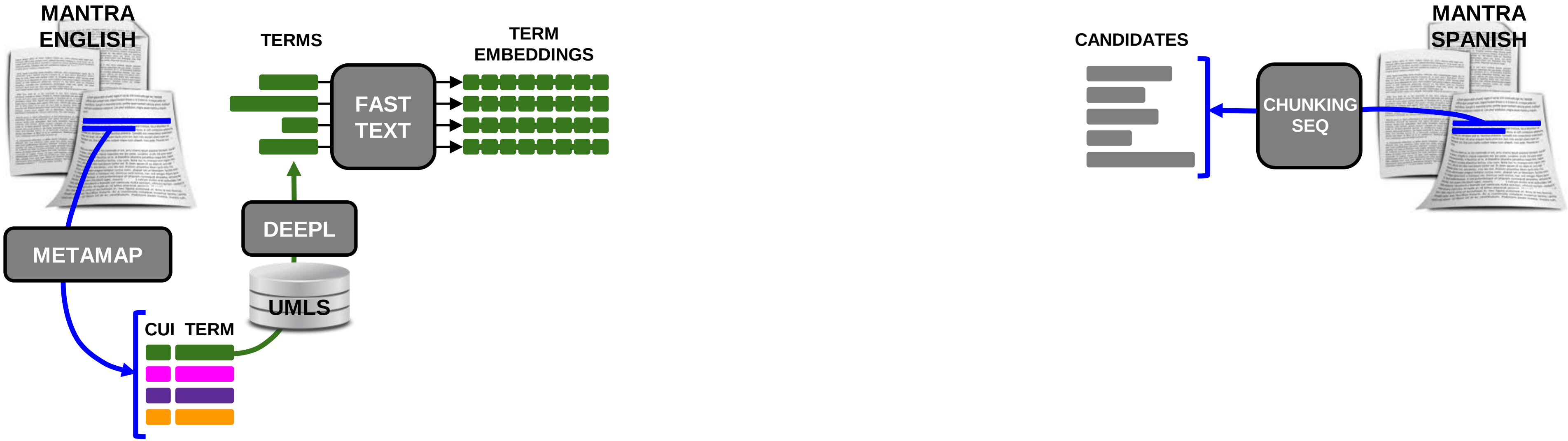
Methods to expand the Spanish terminology



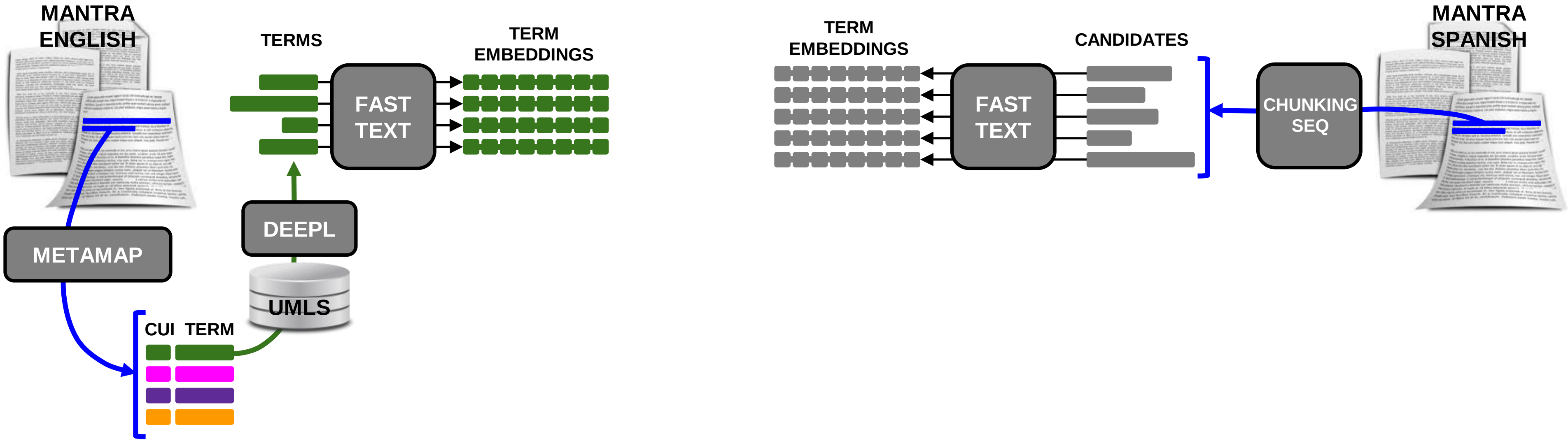
Methods to expand the Spanish terminology



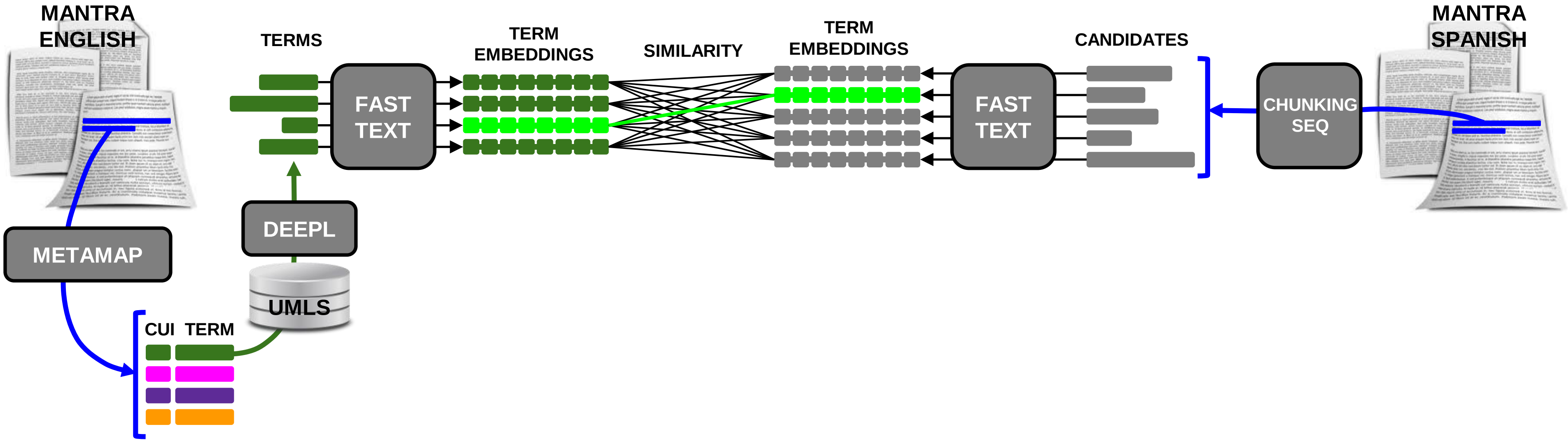
Methods to expand the Spanish terminology



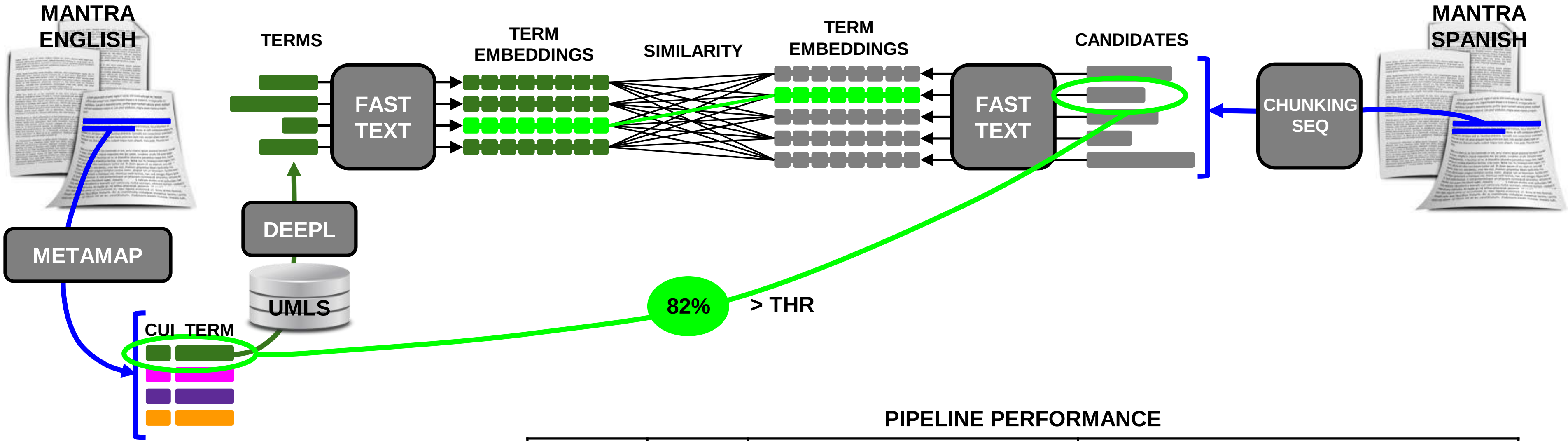
Methods to expand the Spanish terminology



Methods to expand the Spanish terminology



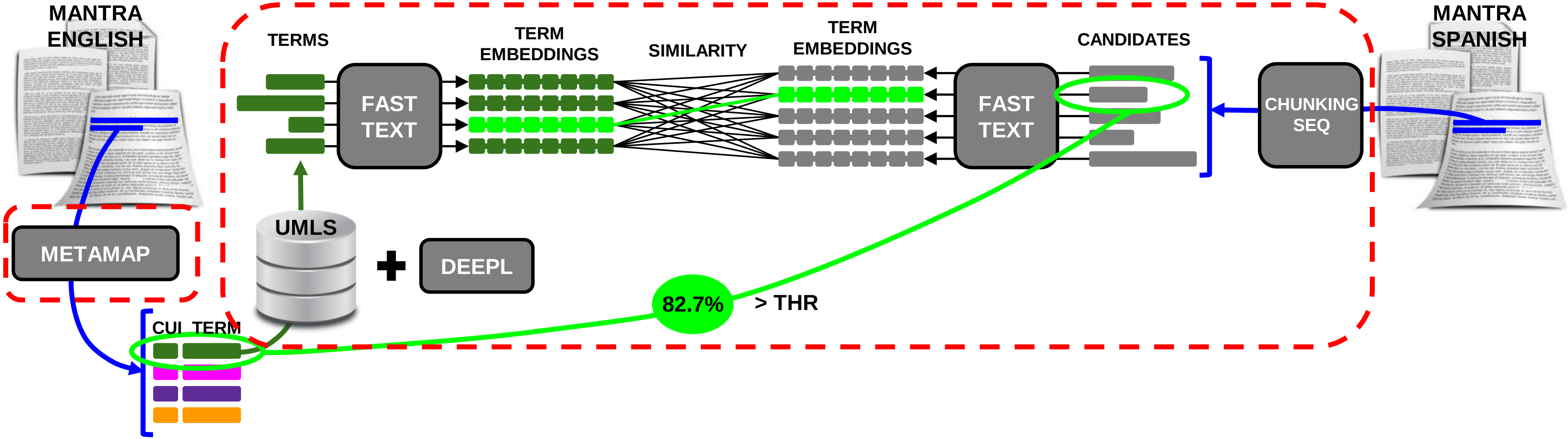
Methods to expand the Spanish terminology



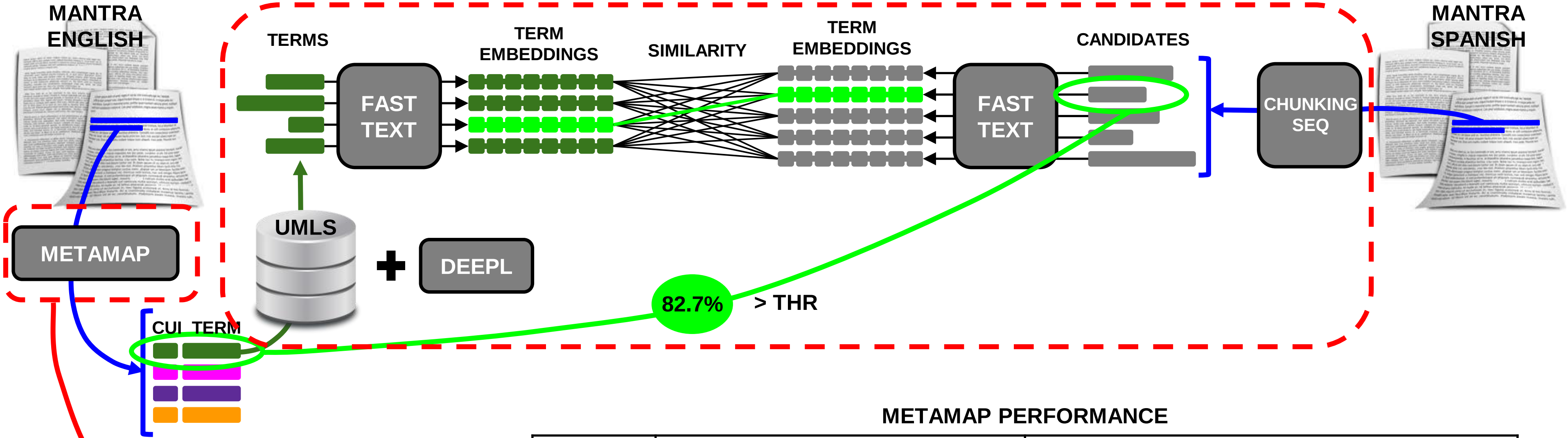
PIPELINE PERFORMANCE

Set	THR	Exact			Overlap			
		P	R	F1	P	R	F1	OP
EMEA	85%	0,817	0,321	0,461	0,846	0,335	0,480	0,989
Medline	82.5%	0,829	0,522	0,643	0,891	0,561	0,689	0,973

Methods to expand the Spanish terminology



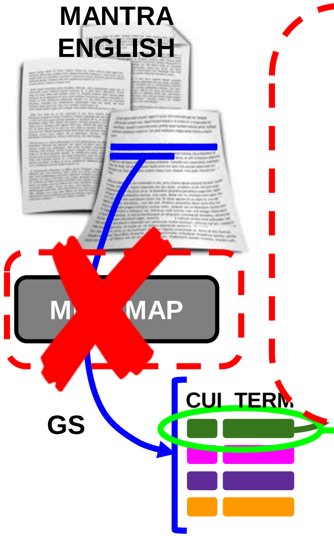
Methods to expand the Spanish terminology



METAMAP PERFORMANCE

Set	Exact			Overlap			
	P	R	F1	P	R	F1	OP
EMEA	0,562	0,387	0,459	0,562	0,387	0,459	1,000
Medline	0,692	0,600	0,640	0,729	0,632	0,677	0,979

Methods to expand the Spanish terminology



TRANSFER PERFORMANCE

Set	THR	Exact			Overlap			
		P	R	F1	P	R	F1	OP
EMEA	85%	0,876	0,683	0,767	0,936	0,748	0,832	0,976
Medline	85%	0,935	0,737	0,824	0,980	0,775	0,865	0,983

PIPELINE PERFORMANCE

Set	THR	Exact			Overlap			
		P	R	F1	P	R	F1	OP
EMEA	85%	0,817	0,321	0,461	0,846	0,335	0,480	0,989
Medline	82.5%	0,829	0,522	0,643	0,891	0,561	0,689	0,973

upf. taln Grup de Recerca en Tractament Automàtic del Llenguatge Natural Automatic Natural Language Processing Research Group

Conclusions

Results of this Study

- Systematic study of tools and resources for NLP in Biomedical & Health domains
- Systematic study of coverage of Spanish Medical Terminologies in comparison to English
- Pipelines for NLP in Spanish and English
- Annotated Parallel Corpora (available to Plan de Impulso)
- Models (e.g. Word Embeddings, Probabilistic Language Models, available to Plan de Impulso)
- Two tested methods of terminology expansion: Direct and Translation
- A series of recommendations and possibilities for implementation
- 10 public deliverables (8 official documents stored at ZENODO)
  - https://zenodo.org/record/3240523

Thanks for your attention

LASTUS Group

- Pablo Accousto
- Ahmed Abura'd
- Kim Cheng Sheang
- Seda Mut
- Horacio Saggion

IXA Group

- Naiara Pérez

Plan TL

- Martin Krallinger
- Mª Ines Rodriguez
- Doaa Ahmed
- David Pérez
- Marta Morales
- Marta Villegas