Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y EMPRESA

Terminologías, ontologías, integración e interoperabilidad de recursos en Sanidad

Martin Krallinger, Antonio Miranda

(BSC-CNS)

mkrallin@bsc.es

Dificultades relacionadas con el lenguaje clínico

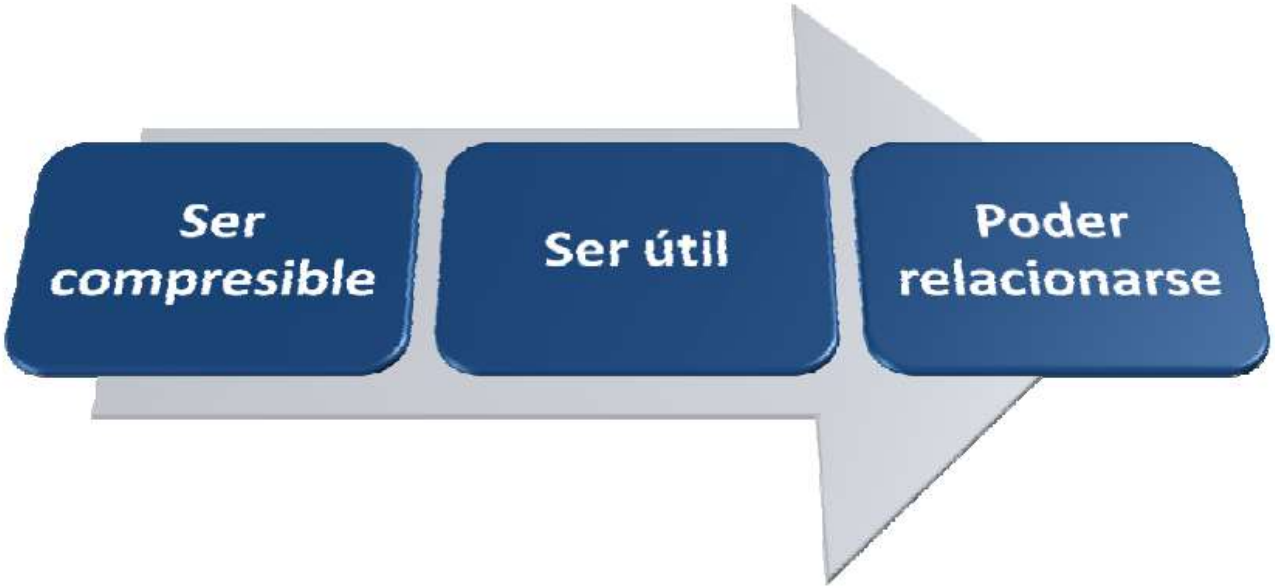
Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Lenguaje usado para la **comunicación del resultado de las observaciones y valoraciones durante el episodio asistencial**

Algunas dificultades:

- Invasión de anglicismos -> proliferación de sinonimia y polisemia
- Uso frecuentes de neologismos
- Uso poco normalizado y excesivo de **abreviaturas, siglas y acrónimos**
- Elipsis, lenguaje **telegráfico** y apócopes (derma en lugar de dermatología)
- Excesivo uso de **epónimos** o términos en el que su significado se encuentra asociado al nombre propio de una persona, lugar, momento, época o cosa (Fiebre del Nilo oeste, trastorno de Asperger)
- Localismos y variantes léxicas derivadas del uso del idioma en una nación, región o zona.
- Pleonasmos (exantema cutáneo)
- Mal uso del género gramatical, el estilo, la **acentuación** y la ortografía,..
- Otros aspectos: **negación**, certeza, frases sin verbo, frases complejas, falta de signos de **puntuación**, falta de acentos, etc.

Aspectos relevantes de recursos terminológicos



1) **Ser comprensible**: significado debe ser comprensible para la comunidad científica internacional

2) **Ser útil**: ser usado en la práctica clínica a nivel internacional. Se deben aportar evidencias de su uso y justificar el beneficio de su disponibilidad en la asistencia

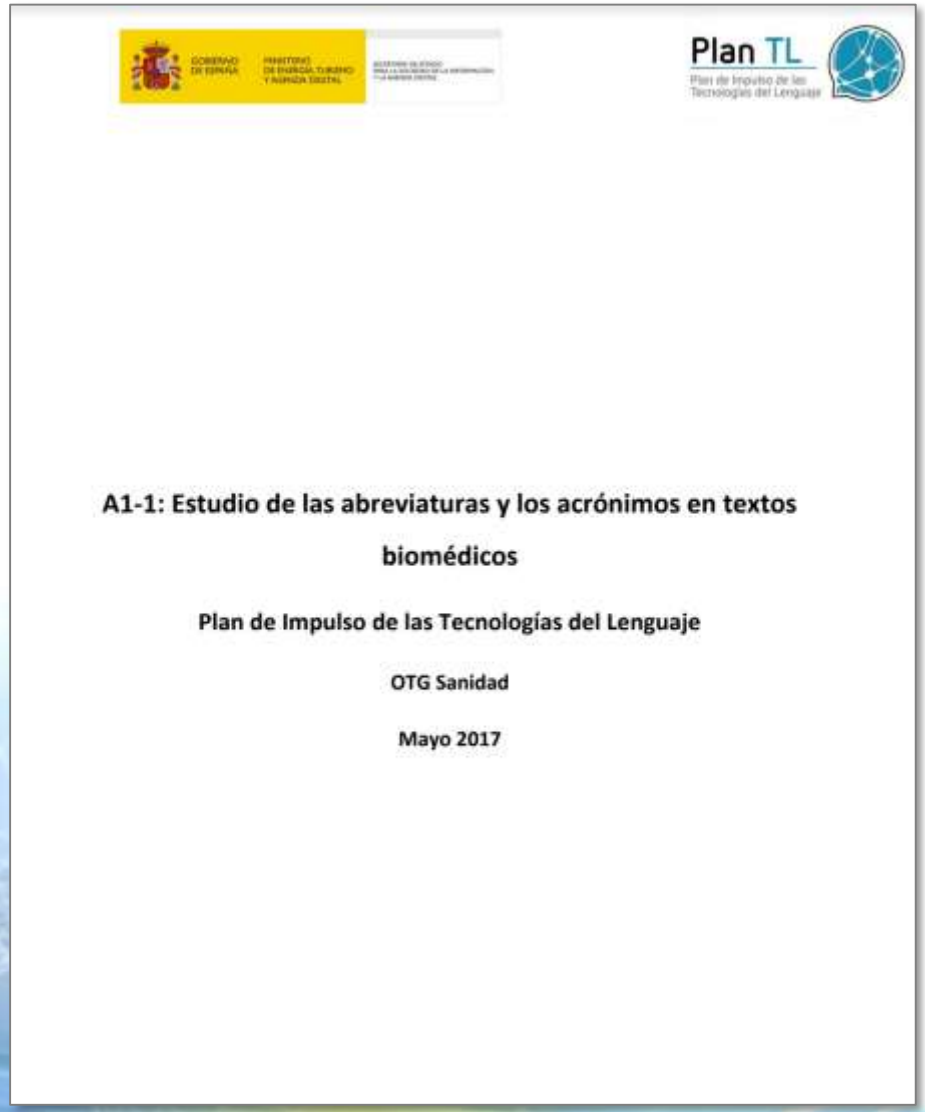
3) **Poder relacionarse**: contenido debe poder relacionarse dentro de la estructura de la terminología clínica.

Interoperabilidad semántica y normalización de la historia clínica electrónica: modelo de producción de activos semánticos basados en estándares. Gonzalo Marco Cuenca (2016)

Estudios Plan TL salud: abreviaturas y ampliación UMLS

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Abreviaturas



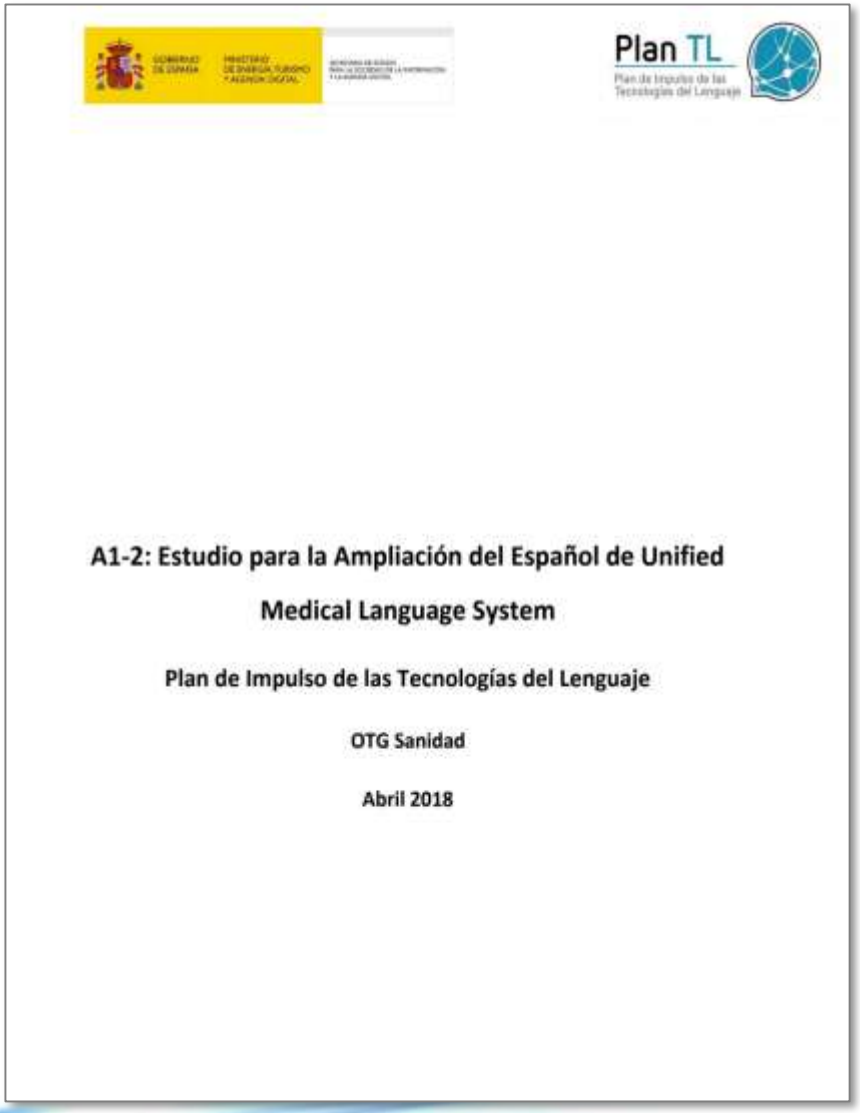
A1-1: Estudio de las abreviaturas y los acrónimos en textos biomédicos

Plan de Impulso de las Tecnologías del Lenguaje

OTG Sanidad

Mayo 2017

Ampliación UMLS



A1-2: Estudio para la Ampliación del Español de Unified Medical Language System

Plan de Impulso de las Tecnologías del Lenguaje

OTG Sanidad

Abril 2018

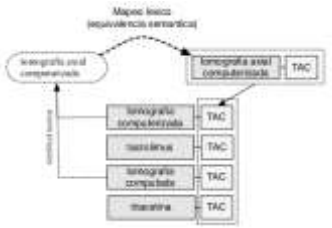
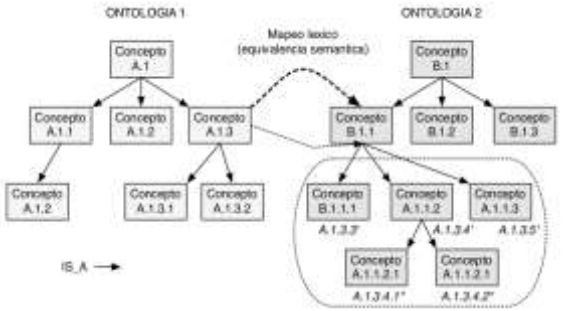
Abreviaturas médicas

- Las abreviaturas usadas en textos clínicos suelen ser **ambiguas** en el 33% de los casos.
- Plasencia y Moliner encontraron casi **22 abreviaturas por informe** clínico (notas de enfermería, informes de alta y altas de emergencia), Benavent et al. en promedio detectaron 14,7 abreviaturas por documento (notas de emergencia, informes de alta y informes de atención especializada).
- **AbreMES-DB**: base de datos de abreviaturas médicas y sus definiciones extraídas automáticamente de SciELO, IBECS y PubMed (https://zenodo.org/communities/medicalnlp)

(34.064 pares abreviatura-definición)

Estudios terminológicos del Plan TL

Cómo paliar la carencia de conceptos:

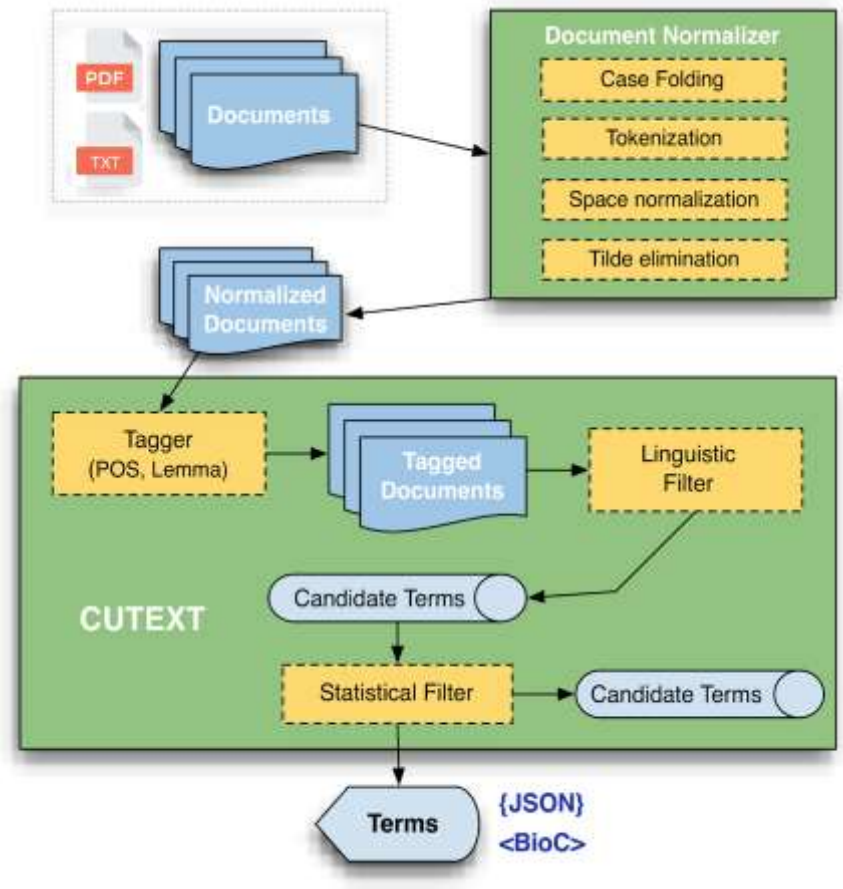
Word embeddings bilingües	Explotar la similitud estructural de los word embeddings en ambos idiomas
Term alignment	 Corpus paralelos
Mapeo léxico	Equivalencia semántica a recursos de abreviaturas y definiciones   Búsqueda de conceptos semánticamente equivalentes

Traducción Manual	Traducir los conceptos en inglés
Traducción Automática	Usar inteligencia artificial para traducir los conceptos en inglés
Glosarios Bilingües	Matching a glosarios bilingües
Reglas morfológicas	A partir de una base común, definir traducciones (v.g. leucocyte, leucocito)
Generación automático de variantes	Mediante reglas. No se obtienen traducciones, se aumenta la cobertura

Para estas estrategias se necesita:

1. **Extracción de términos**.
2. **Mapeo** a terminologías estándar.
3. Recopilación de **documentos paralelos**.

Sistema CUTEXT para extracción de términos



Santamaría, Jesús, and Martin Krallinger. "Construcción de recursos terminológicos médicos para el español: el sistema de extracción de términos CUTEXT y los repositorios de términos biomédicos." Procesamiento del Lenguaje Natural 61 (2018).

Objetivo: extracción automática de términos del dominio usando un sistema simple tipo extractor baseline.

Módulos:

- Conversión al lema.
- Separación de signos y números.
- Cálculo de similitud textual.
- Etc.

Ejemplo:

Además, mientras que la unión constitutiva específica al sitio B peri-kappa se observa en monocitos, la estimulación con ésteres de forbol induce una unión específica adicional.

Enlace:

https://github.com/PlanTL-SANIDAD

Extracción de términos

CuText

- Textos clínicos

Término	Lema	Frecuencia	Idioma
persistencia de sintomatología visual	persistent de sintomatolog visual	2	ES
malaltia d'alzheimer	malalti d'alzheim	3	CAT
patología hemorrágica	patolog hemorrag	1	ES

- Literatura biomédica

Término	Lema	Frecuencia	Área
diabetes mellitus	diabet mellitus	71	Cardio.
trastornos electrolíticos	trastorn electrolit	5	Cardio.
conducto auditivo	conduct audit	413	Anatomía

Extracción de términos de documentos de salud

CuText

Documentos Clínicos

Tipo	# doc.	# tokens / doc	# términos	tokens / término
Casos Clínicos	1 059	1 432.2	164 872	2.3
Gen. Méd. / Enfer. raras	-	-	70 436	2.5
Med. del Trabajo	183	75.1	1 808	1.8
Informes de alta	3 348	1 217.8	258 737	2.2
Corpus Salud*	-	-	373 399	2.2

Libros salud/medicina

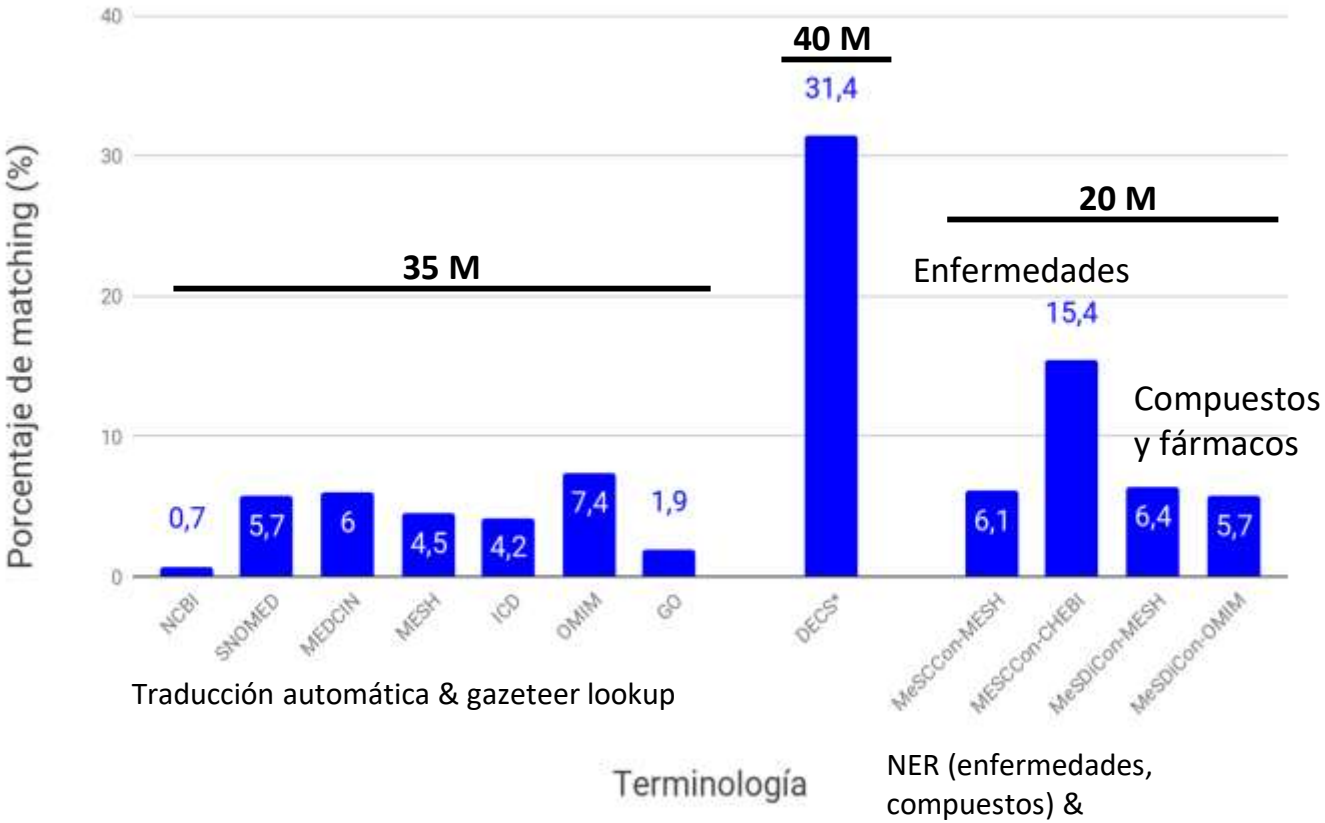
Área médica	# doc.	tokens / doc	# términos	tokens / término
Cirugía	24	217 934.4	306 702	2.2
Anatomía	9	181 300.6	113 798	2.1
Cardiología	7	32 762.7	21 508	2.0
Radiología	10	105 935.1	170 266	2.2

*Recursos corpus salud: casos clínicos, artículos medicina, informes de medicamentos,..

Detección de términos en HCE

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Porcentaje de detección de términos en el corpus clínico de Atención Primaria de las 7 terminologías con mayor número de términos en UMLS, de DeCS y de otros recursos generados



Traducción automática & gazeteer lookup

NER (enfermedades, compuestos) & Traducción automática & gazeteer lookup

Terminología	# términos	# mapeos
NCBI	1,6M	10 705
SNOMED	1,2M	68 129
MEDCIN	860K	51 468
MESH	810k	36 683
ICD	650k	27 501
OMIM	170k	12 388
GO	160k	3 019
DECS*	60k	19 611

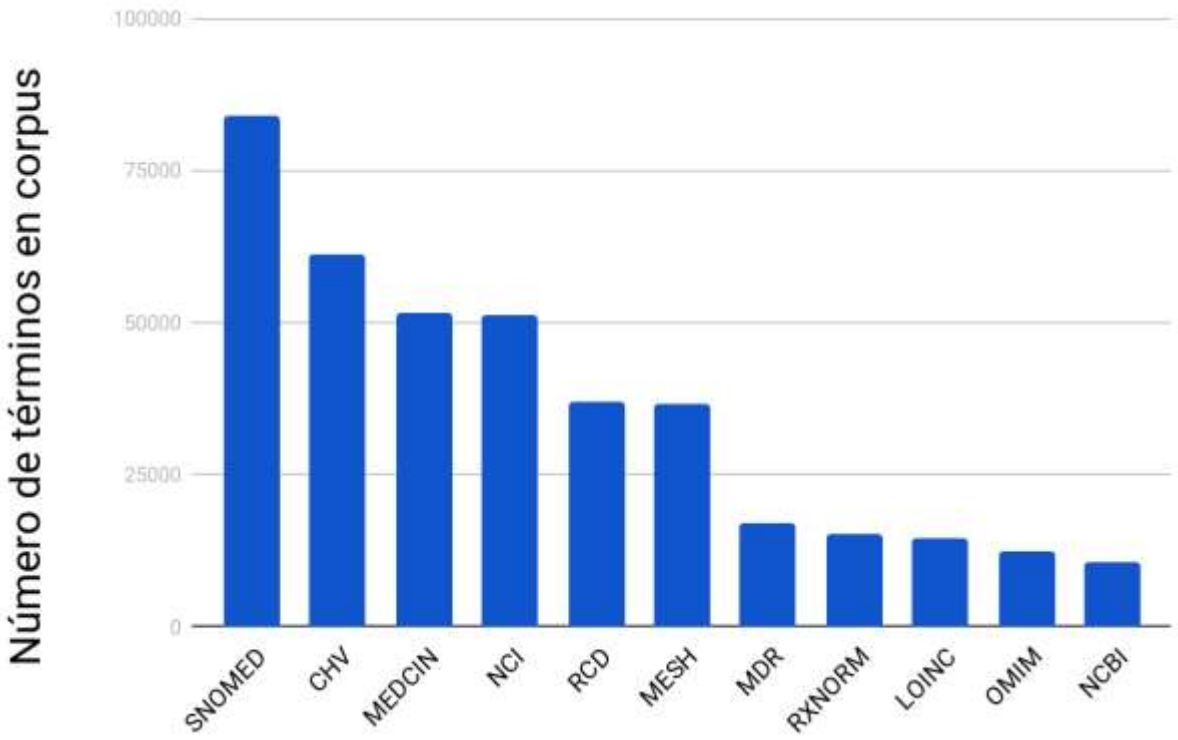
Recursos mapeados	# términos	# mapeos
MeSCCon - MESH	560k	**33 909**
MESCCon - CHEBI	70k	**33 909**
MeSDiCon - MESH	3.5M	**222 701**
MeSDiCon - OMIM	30k	**1 942**

UMLS: Unified Medical Language System
DeCS: Descriptores en Ciencias de la Salud

Detección de términos en HCE

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Número de términos detectados en corpus clínico de Atención Primaria: top 12 terminologías



Terminología	Descripción
SNOMED	Systematized Nomenclature of Medicine
CHV	Consumer Health Vocabulary
MEDCIN	For electronic medical record
NCI	National Cancer Institute Thesaurus
RCD	Read Codes - care and treatment
MESH	Medical Subject Headings
MDR	Medical Dictionary for Regulatory Activities
RXNORM	Clinical drugs for humans
LOINC	Logical Observation Identifiers Names and Codes
OMIM	Online Mendelian Inheritance in Man
NCBI	NCBI terminology

Estadísticas generales del mapeo:

- Terminologías probadas: 129
- Terminologías con algún término en corpus: 128
- Términos únicos: 8 906 988
- Términos únicos que hacen matching: **561 896**

Tareas terminológicas claves

	Extracción	Mapeo	Traducción	Indización
Relevancia	Punto de partida	Normalización e interoperabilidad	Generación de nuevos datos, extracción de información, etc.	Mejora la extracción de información
Dificultades	Lenguaje, Convenciones, terminologías, etc.	Lenguaje, Convenciones, terminologías, alcance de las terminologías, etc.	Ambigüedades, validación, etc.	Ausencia de corpus, vocabularios muy extensos, jerarquías con relaciones complejas, etc.
Recursos generados zenodo.org/communities/medicalnlp	CuText PharmaCoNER AbreMES-DB	HPO SNOMED CIE10	MeSpEn	CodiEsp MESINESP
Campañas temu.bsc.es	eHealth-KD* knowledge-learning.github.io/ehealthkd-2019/	MedTermMap	WMT19 statmt.org/wmt19/biomedical-translation-task.html MedTrans	MESINESP temu.bsc.es/mesinesp/ CodiEsp temu.bsc.es/codiesp/
	Validación			

Tareas de mapeo de términos

Recursos:

2 tipos de recursos de mapeo de términos:

- **Mapeos** de términos a un vocabulario de referencia (mapping):

Término candidato	Código
absceso de la mama	N611
cardiopatía isquémica	I259
Complicaciones del tratamiento endovascular	T801XXA

- **Clasificación** de parejas término candidato-término de referencia según su relación (clasificación):

Candidato	Término en ontología	Código	Relación
lesión traumática de arteria braquial	Traumatismo de la arteria braquial	S451	Exact
hematoma de glándula tiroides	Trastornos de la glándula tiroides	|IV1	Narrow-to-broad
Insuficiencia renal y complicaciones vasculares	Insuficiencia renal terminal	N180	Broad-to-narrow

Recurso Gold Standard manual para sistemas de mapeo de términos

Recursos:

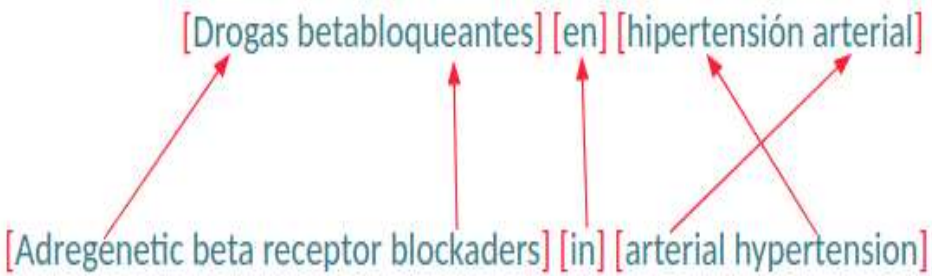
	HPO		CIE10		SNOMED	
Tipo de diccionario	**Mapping**	**Clasificación**	**Mapping**	**Clasificación**	**Mapping**	**Clasificación**
# términos	1 988	4 989	1 000	5 000	2 000	5 000
% términos relacionados	21%	Exacto - 0.6% N-to-B - 1% B-to-N - 2%	88%	Exacto - 11% N-to-B - 5% B-to-N - 12%	79%	Exacto - 7% N-to-B - 7% B-to-N - 12%
Campañas relacionadas*	**2019 n2c2. Track 3 (**n2c2.dbmi.hms.harvard.edu/track3) **ShARe/CLEF eHealth 2013 Task 1** (sites.google.com/site/shareclefehealth/task-description)					

Campañas:

- **MedTermMap** (2020)

MeSpEn

- Documentos Paralelos: reconocimiento de términos o entidades nombradas en inglés y reconocimiento de términos candidato en español mediante alineamiento de frases (temu.bsc.es/mespen/)



- Glosarios Bilingües (zenodo.org/record/2205690#.Xd_HWtEo9hF)

Inglés-Español	fibrin degradation products	productos de degradación de la fibrina
Francés-Español	réactivité face à la nouveauté	reacción ante la novedad
Alemán-Ruso	Bakterienkultur	бактериологический посев
etc		

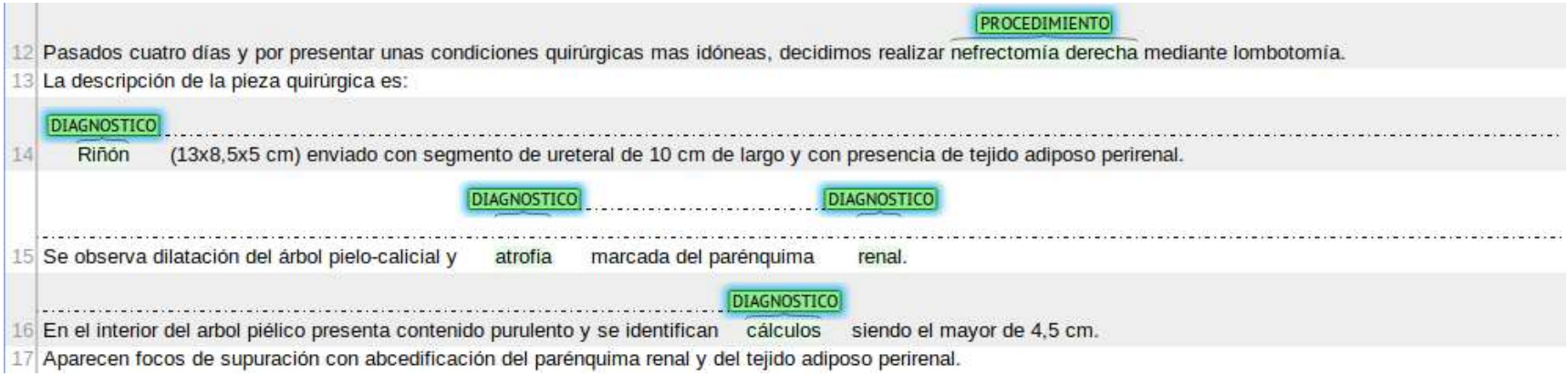
MespEN glosarios médicos bilingües

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Language pair	Frequency	Language pair	Frequency	Language pair	Frequency
English-Spanish	123,788	Latin-Russian	2,486	German-Russian	225
English-Korean	69,368	German-Swedish	2,208	English-Romanian	205
Chinese-English	66,939	German-Portuguese	2,028	Italian-Spanish	196
English-Italian	24,155	English-Swedish	1,067	Danish-English	193
English-German	18,534	German-Italian	976	French-Spanish	179
English-Japanese	18,320	English-Slovenian	945	Danish-Polish	166
Arabic-English	9,384	Bengali-English	841	Russian-Spanish	122
English-Turkish	7,675	English-Thai	835	English-Hindy	120
German-Spanish	7,004	Dutch-French	585	French-German	119
Dutch-English	6,878	English-Indonesian	491	French-Italian	117
English-French	6,571	Bulgarian-English	347	Croatian-German	115
English-Russian	4,346	Croatian-English	339	German-Romanian	109
English-Polish	3,727	Polish-Spanish	271	Dutch-Spanish	70
English-Hungarian	2,711	Dutch-Turkish	238	Portuguese-Spanish	61
English-Greek	2,626	Latin-Polish	237	English-Norwegian	44
English-Portuguese	2,517	Croatian-French	235	**TOTAL**	**390,713**

Generados por 500 traductores profesionales.

Indización/codificación documentos: CodiEsp (CIE10)



Codificación:

Mención	Tipo de mención	CIE-10	Término CIE-10
nefrectomía derecha	PROCEDIMIENTO	0tt0	Resección Riñón Derecho
atrofia renal	DIAGNOSTICO	n26.1	Atrofia del riñón
Riñón cálculos	DIAGNOSTICO	n20.0	Cálculo del riñón

Indización de documentos: Mesinesp (descriptores DeCS)

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Recursos:

	MESINESP	CodiEsp
# documentos	318 658	1 000
Tipo de documento	resumen de artículo	caso clínico
# caracteres / documento	1141	2337
# códigos / documento	8	21
Códigos	DeCS	CIE10
Tipo de indización	En la base de datos	Por expertos

Campañas:

- **MESINESP** en ECIR 2020 (temu.bsc.es/mesinesp/)
- **CodiEsp** en MIE 2020 (temu.bsc.es/codiesp/)

Gracias

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

- Martin Krallinger
- Marta Villegas
- Siamak Barzegar
- Antonio Miranda
- Alejandro Asensio
- Aitor Gonzalez
- Montse Marimon
- Felipe Soares

- Alfonso Valencia (BSC Life)

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

David Perez (SEAD)

- *AQuAS* (Miguel Gallofre López)
- *AEMPS-BIFAP* (Julio Bonis Sanz)
- AEMPS-FTM (JM Simarro)
- *FID-Salud/MSSSI* (Elena García)
- *FISEVI/Hosp. Virgen del Rocio* (Carlos Parra)
- *Hospital 12 de Octubre* (Pablo Serrano)
- *IBECS/Carlos III* (Elena Primo)
- *Informática Médica Hosp. Clínic* (Raimundo Lozano)
- *MSSSI* (Maribel García Fajardo)
- *RANM* (Cristina V. González)

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

cima
CENTER FOR APPLIED MEDICAL RESEARCH
UNIVERSITY OF NAVARRA

- Obdulia Rabal
- Julen Oyarzabal

BIO CREATIVE

- BioCreative organizers
- Cecilia Arighi/Cathy Wu (Uni. Delaware)
- Lynette Hirschman (MITRE)

Universidade Vigo

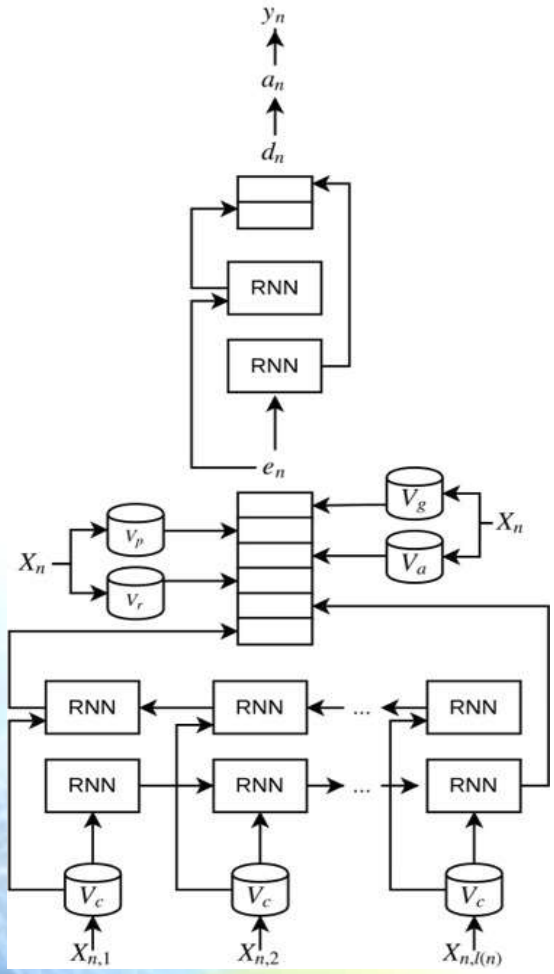
- Analia Lourenço
- Martin Perez Perez
- Gael Perez Rodriguez
- Florentino Fernández Riverola

Plan TL

Plan de Impulso de las Tecnologías del Lenguaje

Extracción y clasificación de medicamentos

PharmaCoNER Tagger



Objetivo: extracción de términos automáticamente de textos biomédicos en español y clasificación (NER).

Core: AI system

Módulos:

- Diccionario de términos del dominio
- Diccionario de sufijos
- Embeddings con información morfosintáctica

Campaña: PharmaCoNER en EMNLP-IJCNLP 2019 (Hong Kong)

Enlace:
github.com/TeMU-BSC/PharmaCoNER-Tagger

Recursos terminológicos en PLN y su uso clínico

En PLN:

- **Extracción** de información (v. g.: RE)
- **Traducción automática**
- **Tareas: Codificación** de documentos
- Reconocimiento de entidades nombradas

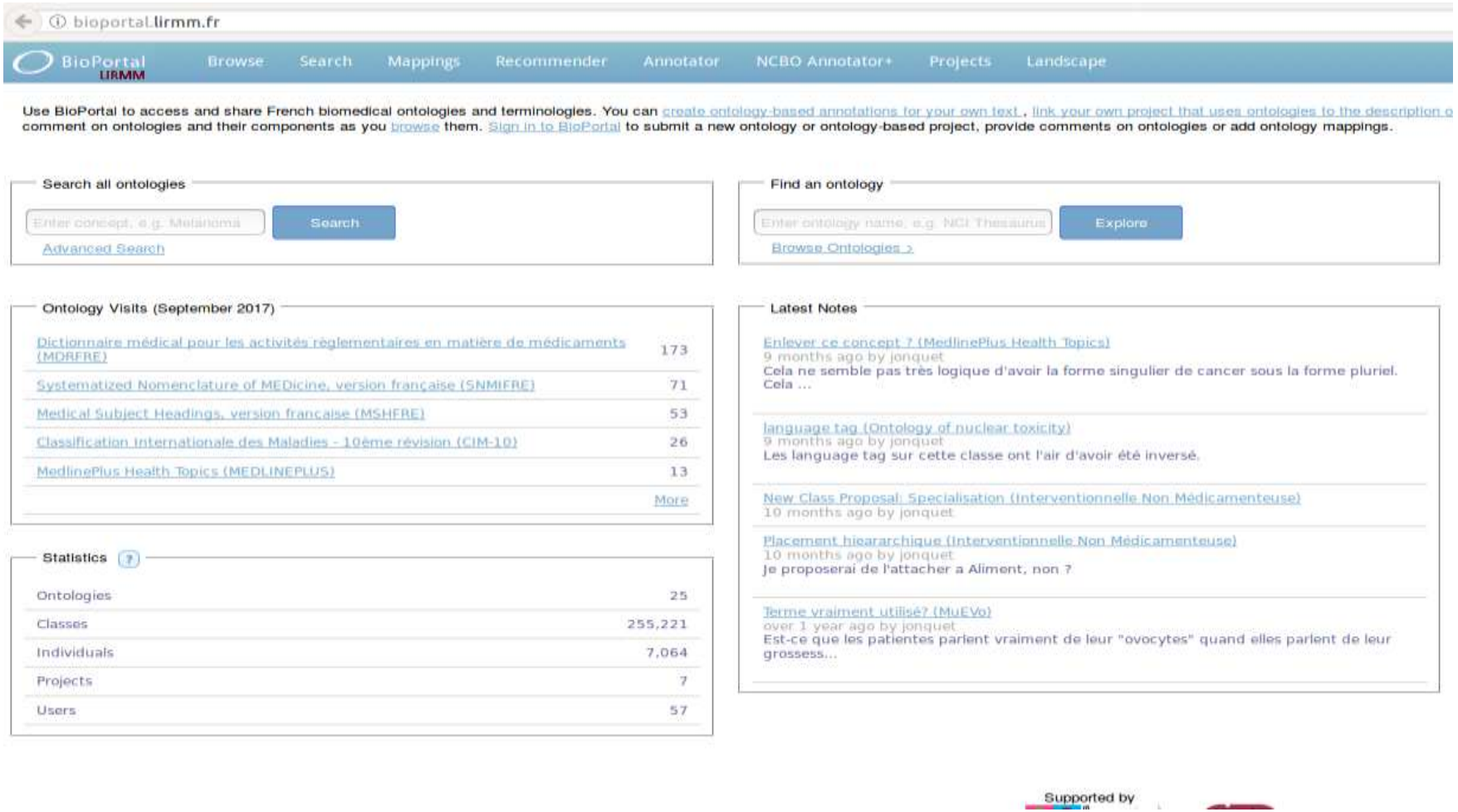
En salud:

- **Describir** las condiciones de los pacientes con precisión
- **Coordinar** los departamentos de un hospital
- Permitir la **comunicación** entre el personal sanitario con las aseguradoras y las farmacias

Necesidades de los textos biomédicos y clínicos:

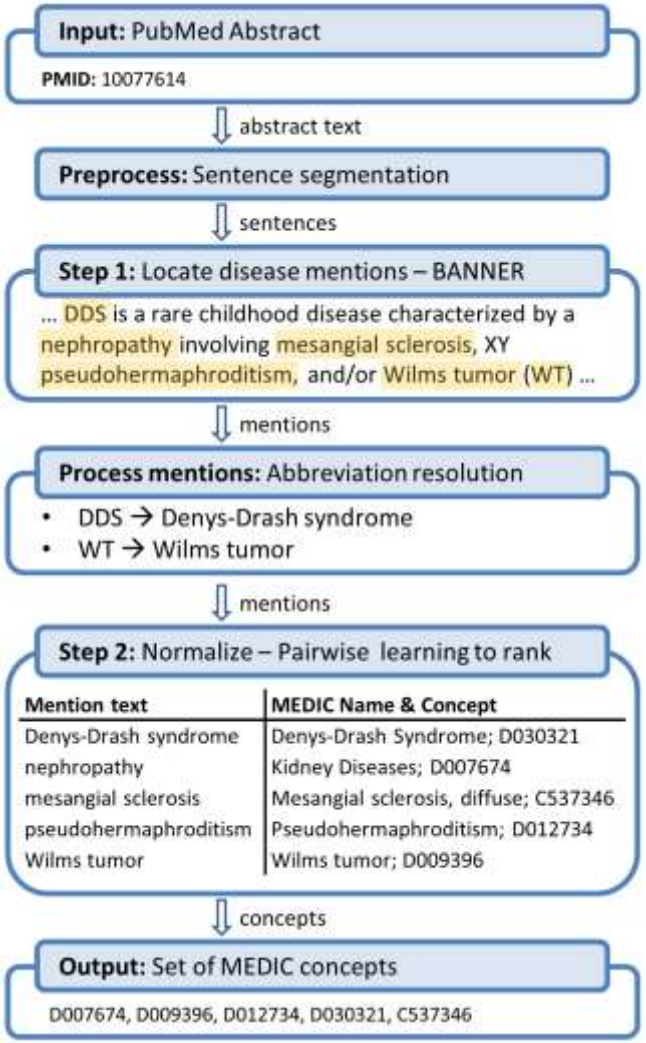
Términos técnicos, abreviaturas, errores tipográficos, frases complejas, falta de signos de puntuación, etc.

1. BioPortal


bioportal.lirmm.fr
BioPortal
LIRMM
Browse
Search
Mappings
Recommender
Annotator
NCBO Annotator+
Projects
Landscape
Use BioPortal to access and share French biomedical ontologies and terminologies. You can create ontology-based annotations for your own text , link your own project that uses ontologies to the description o
comment on ontologies and their components as you browse them. Sign in to BioPortal to submit a new ontology or ontology-based project, provide comments on ontologies or add ontology mappings.
Search all ontologies
Enter concept, e.g. Melanoma
Search
Advanced Search
Find an ontology
Enter ontology name, e.g. NCI Thesaurus
Explore
Browse Ontologies >
Ontology Visits (September 2017)
Dictionnaire médical pour les activités réglementaires en matière de médicaments (MORFRE) 173
Systematized Nomenclature of MEDicine, version française (SNMIFRE) 71
Medical Subject Headings, version française (MSHFRE) 53
Classification Internationale des Maladies - 10ème révision (CIM-10) 26
MedlinePlus Health Topics (MEDLINEPLUS) 13
More
Latest Notes
Enlever ce concept ? (MedlinePlus Health Topics)
9 months ago by jonquet
Cela ne semble pas très logique d'avoir la forme singulier de cancer sous la forme pluriel. Cela ...
language tag (Ontology of nuclear toxicity)
9 months ago by jonquet
Les language tag sur cette classe ont l'air d'avoir été inversé.
New Class Proposal: Specialisation (Interventionnelle Non Médicamenteuse)
10 months ago by jonquet
Placement hiearchique (Interventionnelle Non Médicamenteuse)
10 months ago by jonquet
Je proposerai de l'attacher a Aliment, non ?
Terme vraiment utilisé? (MuEVo)
over 1 year ago by jonquet
Est-ce que les patientes parlent vraiment de leur "ovocytes" quand elles parlent de leur grossess...
Statistics
Ontologies 25
Classes 255,221
Individuals 7,064
Projects 7
Users 57
Supported by

1. Objetivo Final

DNorm: Disease Named Entity Recognition and Normalization with Pairwise Learning to Rank



1. Objetivo

recopilación de recursos lingüísticos

- 42 recursos
- 386 elementos
- 8,5 GB

1. Objetivo

recopilación de recursos lingüísticos

- 42 recursos
- 386 elementos
- 8,5 GB

2. Definición

Término: es una unidad léxica que designa a un concepto en un campo temático particular

Sager, J.: *Pour une approche fonctionnelle de la terminologie*. In: Bènjoint, H; Thoiron, P. (Ed.). Le sens en terminologie. Lyon: Presses Universitaires de Lyon, (2000) 40-60

Marincovich, J.: *Palabra y término: ¿Diferenciación o complementación?*. Revista Signos: Estudios de Lingüística, Valparaíso, v.41, n.67, (2008) 119-126

2. Definición

Término Médico: Sintagma, generalmente nominal, que posee significado para el área médica y al que es posible acceder a través de un proceso de detección automática basado en información lingüística porque:

I.Su lema es una entrada léxica en un diccionario electrónico del dominio médico.
II.Posee una estructura morfológica propia del dominio médico que se puede formalizar y ser implantada en una máquina.
III.Incluye un neologismo que no posee una estructura morfológica, pero su categoría gramatical puede ser deducida automáticamente a través del contexto sintáctico.

3. Utilidades

Punto de partida para realizar tareas más complejas

Elaboración de listas de entradas para diccionarios especializados

Creación de base de datos

Creación de ontologías

Creación de taxonomías

...

4. Inconvenientes y Dificultades

Cambio constante en la terminología: necesidad de herramientas que puedan detectar los términos nuevos, así como las variaciones.

Las tareas de extracción suelen ser específicas: deben adaptarse a los requerimientos y particularidades propias de cada una de ellas.

Variaciones léxicas, sinonimia, homonimia.

Falta de convenciones firmes en la nomenclatura: existen directrices pero no se imponen restricciones; junto con los términos "bien formados" existen otros ad-hoc, que son problemáticos para los sistemas de identificación de términos.

Krauthammer, M. & Nenadic, G. Term identification in the biomedical literature. Journal of Biomedical Informatics, San Diego, v.37, n.6, (2004) 512-526.

4. Algunas Propuestas

Sistemas basados en:
1.Características internas: ortografía (mayúsculas, dígitos, caracteres griegos, etc.)
2.Pistas morfológicas (afijos específicos y formantes cultos, principalmente griegos y latinos)
3.Información procedente del análisis sintáctico
4.Medidas estadísticas para promover candidatos a términos

5. Antecedentes – TerMine

Basado en Termine. Combina combina **dos filtros**: uno **lingüístico** y uno **estadístico**.

El **filtro lingüístico** basado en **expresiones regulares**, y una **stop-list**.

El **análisis estadístico** se basa en:
1) La frecuencia de ocurrencia del término candidato
2) La frecuencia del término candidato como parte de otros términos candidatos más largos
3) El número de estos términos candidatos más largos
4) La longitud del término candidato

Frantzi, K., Ananiadou, S. and Mima, H. (2000) Automatic recognition of multi-word terms. International Journal of Digital Libraries 3(2), pp.117-132.

6. CUTEXT

Implementación: Java

Idiomas: inglés, castellano, catalán, gallego. Pruebas de evaluación para inglés y castellano

Etiquetadores: TreeTagger, GeniaTagger

Documentos: teclado(texto), fichero/s (pdf, texto)

6. CUTEXT – Texto

Fichero Texto Plano

Fichero Formato JSON

Fichero Formato BioC (XML)

6. CUTEXT – JSON


```
{
        "terms_ENGLISH_genia-corpus":
        [
                {
                        "term": "T cells",
                        "frecuency": "1087",
                        "c-value": "3291.7436502440028"
                },
                {
                        "term": "NF-kappa B",
                        "frecuency": "530",
                        "c-value": "1674.6077201257863"
                },
                {
                        "term": "transcription factors",
                        "frecuency": "350",
                        "c-value": "1067.1730638057595"
                },
```


Términos ordenados por su c-value

6. CUTEXT - BioC

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE collection SYSTEM "BioC.dtd">
<collection>
  <source>Unknown</source>
  <date>20170907</date>
  <key>nothing.key</key>
  <document>
    <id>1</id>
    <passage>
      <infon key="type">paragraph</infon>
      <offset>0</offset>
      <text>Further, while specific constitutive binding to the peri-kappa B site
is seen in monocytes, stimulation with phorbol esters induces additional,
specific binding. Understanding the monocyte-specific function of the peri-
kappa B factor may ultimately provide insight into the different role
monocytes and T-cells play in HIV pathogenesis.</text>
      <annotation id = "0">
        <infon key = "type">Term</infon>
        <infon key = "cvalue">112.46797000057693</infon>
        <infon key = "frecuency">1</infon>
        <location offset = "15" length = "29" />
        <text>specific constitutive binding</text>
      </annotation>
```

Información de los Términos

6. CUTEXT – Serializada

Se serializa el resultado final:

- Se almacenan los términos extraídos por TermineSP con todos sus atributos: término, frecuencia, c-value, etc.

- También se almacenan los términos extraídos por aquellos corpus que los posean, es decir los que estén anotados (por ejemplo, el corpus Genia).

7. Sistemas por Casos de Uso

Enfocados en el español:

Castro, E.; Iglesias, A.; Martínez, P.; Castaño, L.: *Automatic identification of biomedical concepts in Spanish language unstructured clinical texts*. In: CASTRO, E. et al. In: AC INTERNATIONAL HEALTH INFORMATICS SYMPOSIUM, 1., 2010, Nueva York. Proceedings, Nueva York: ACM (2010) 751-757.

Detección de conceptos de notas clínicas y su asociación en la ontología SNOMED-CT.

Sánchez, D.; Batet, M. & Valls, A.: *Web-based semantic similarity: an evaluation in the biomedical domain*. Int. J. Software and Informatics, Beijing, v.4, n.1, (2010) 39-52

Garla, V. & Brandt, C. *Semantic similarity in the biomedical domain: an evaluation across knowledge sources*. BMC Bioinformatic 2012, Londres, v.13, n.261, (2012).

Analizar automáticamente la relación entre conceptos que comparten el mismo contexto.

7. Sistemas por Casos de Uso

Enfocados en el español:

Vivaldi, J. & Rodríguez, H.: *Using Wikipedia for term extraction in the biomedical domain: first experiences*. Procesamiento de Lenguaje Natural, Jaén, v. 45, (2010) 251-254

Sistema de extracción de términos probado en un corpus médico. Encontrado un candidato a término, los pasos son: (1) encontrar una página de Wikipedia que se corresponda con el candidato, (2) encontrar todas las categorías de Wikipedia asociadas a dicha página, y por último, (3) explorar la Wikipedia siguiendo recursivamente todos los enlaces de categorías encontrados en (2) a fin de enriquecer la frontera del dominio.

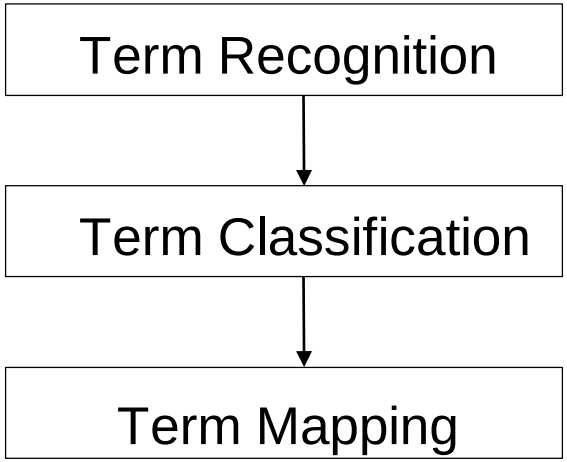
Moreno-Sandoval, A.; Campillos-Llanos, L. *Desing an annotation of multimedica: a multilingual text corpus of the biomedical domain*. Procedia: Social and Behavioral Sciences, Amsterdam, v.95, (2013) 33-39

Elaboración de un corpus compuesto por textos biomédicos en español, árabe y japonés.

8. Trabajo Futuro

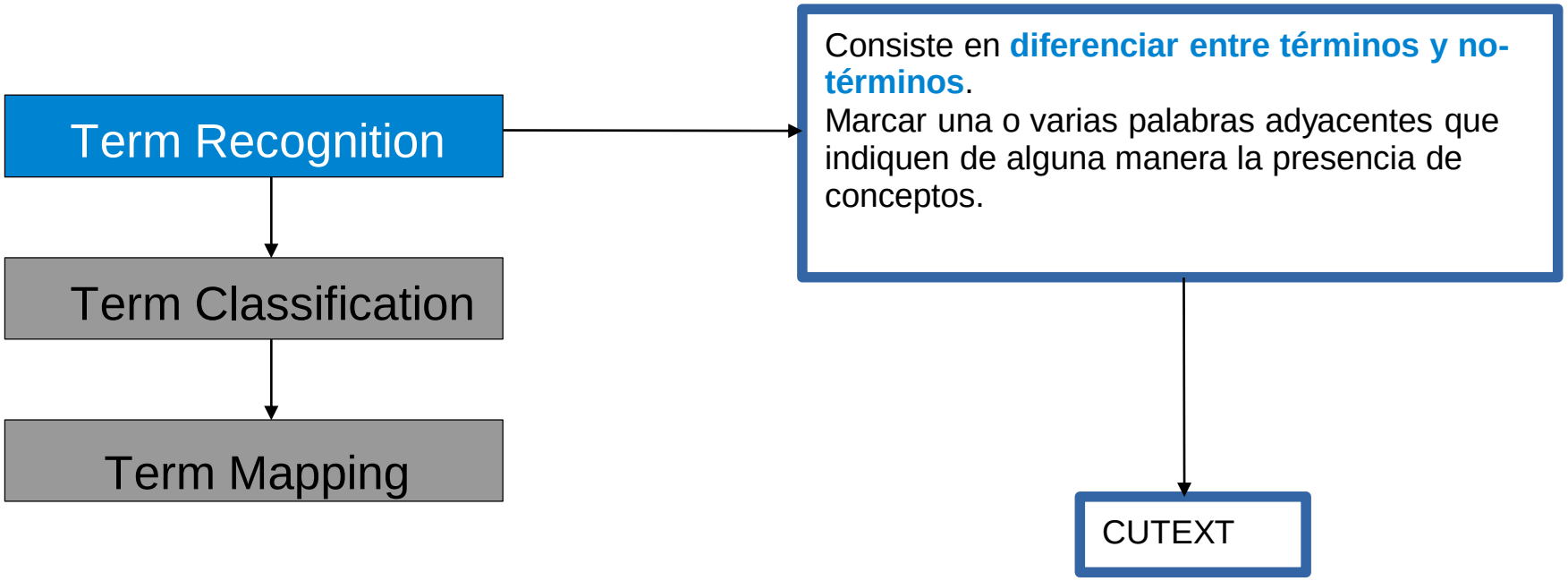
Según Krauthammer y Nenadić[1], para que la **identificación de términos** tenga éxito se deben realizar **3 etapas** secuenciales:

1)Reconocimiento del término
2)Clasificación del término
3)Emparejamiento del término

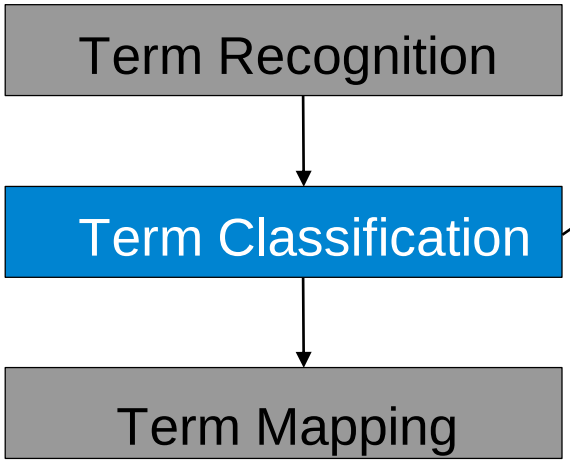


1 Michael Krauthammer, Goran Nenadić. Term Identification in the Biomedical Literature.

8. Trabajo Futuro



8. Trabajo Futuro

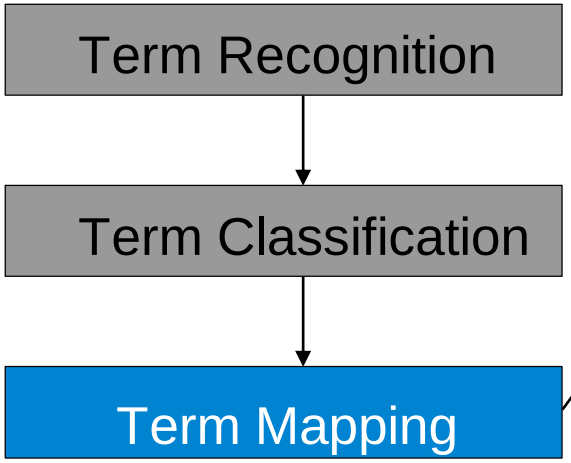


Consiste en **asignar los términos al dominio específico**. Es decir, quedarse sólo con los términos pertenecientes al dominio.
Es una etapa necesaria para las aplicaciones que trabajan con términos específicos (médicos, biológicos, biomédicos, etc.).

Técnica Estadística

Objetivo: medir el grado de distintividad de un término en un corpus especializado en contraste con su frecuencia en un corpus general.
Métricas más empleadas: log-likelihood ratio test, logDice
La **idea** central: qué términos son sobreutilizados o infrautilizados en nuestro corpus de análisis en comparación con la frecuencia de las mismas palabras en un corpus de referencia.

8. Trabajo Futuro



Vincula los **términos con conceptos** bien definidos de fuentes de datos referentes, como vocabularios controlados o bases de datos.
Los términos asignados se anotan con identificadores de referencia (IDs) que actúan como claves para la información suplementaria, como términos preferidos y sinónimos, o información de secuencia.
El mapeo de términos es esencial en cualquier esfuerzo de integración de datos.

UMLS (Unified Medical Language System)

Incluye tesauros y terminologías, como: Medical Subject Headings (MeSH), Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT) o la versión 10 de la Clasificación Internacional de Enfermedades (ICD-10).
Tiene códigos identificadores unicos de concepto (**CUI**) asociados a cada variante terminológica en los distintos recursos. Por ejemplo, el código C0817096 designa al término *pecho* o *caja torácica* en MeSH, y también al término *torácico* o *tórax* en SNOMED-CT.

Comparison between biomedical and clinical text mining and natural language processing (1)

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Aspects	Biomedical/Biology	Medical/clinical
Documents & datasets	• Scientific literature • Free text in annotated databases • Medicinal chemistry patents	• Electronic health records • Medical literature • Product/drug labels • Clinical trials • Medicines Agency reports • Social media
Main entities of interest	• Genes/proteins • Chemical compounds/drugs • Diseases • Organisms & species • Cell types/cell lines • Sequence variants/mutations	• Diseases • Treatments • Drugs • Signs/symptoms • Anatomical locations • Diagnosis • Adverse events
Relations, events	• Protein-protein interaction • Drug-disease • Gene-Disease • Gene/protein-Gene Ontology • Gene regulation • Drug-drug interaction	• Disease-treatment • Drug-adverse event • Spatial relations • Temporal relation • Disease-disease relation • Sentential negation

Comparison between biomedical and clinical text mining and natural language processing (2)

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Aspects	Biomedical/Biology	Medical/clinical
Language	• Predominantly English	• Any language with EHRs
Difficulties & challenges	• Long complex sentences • Gene/protein mention grounding • Technical terminology • Use of abbreviations • Relevant information not running text (figures, tables)	• Heavy use of abbreviations • Telegraphic writing • Ungrammatical sentences • Lack of punctuation marks • Uncertainty, negation • Temporality • Misspellings
Access issues	• Legal: Copyright restrictions (full text) • Technical: document types, formats, additional materials	• Legal: Confidentiality, Privacy (de-identification and anonymization) • Technical: structured & unstructured narrative text

Linguistic infrastructures generated by Plan TL (1)

https://github.com/PlanTL-SANIDAD

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

	Description https://zenodo.org/communities/medicalnlp
Terminological resources	• Registry of linguistic / **terminological resources** (368 in total, 103 of medical domain) • Spanish medical **abbreviation** definition database (> 34,064 , AbreMES) • Spanish **medical term database** generated by CUTEXT – automatic term recognition of the medical domain in Spanish from the medical literature, EHRs, etc. • **Bilingual medical glossary** of for multiple language pairs (MeSpEN <400 thousand) • Spanish medical controlled vocabularies, ontologies and medical entity gazetteers generated by deep learning based **medical machine translation** • Medical Word **embeddings** generated from Spanish medical corpora
Medical concept recognition and semantic tagging	• **Gazetteer-lookup**: System of recognition of terms from a terminology (tested with Snomed, adaptation of GATE • **Automatic Term recognition** (ATR): a system called CUTEXT for the recognition of terms using linguistic-statistical standards • PharmaCONER Tagger: **deep learning based system for detecting drugs, chemical compounds** and genes in clinical texts • **Time expression** entity recognition system HeidelTime grammar for temporal tagging of Spanish Electronic Health Records

Linguistic infrastructures generated by Plan TL (2)

https://github.com/PlanTL-SANIDAD

	Description https://zenodo.org/communities/medicalnlp
Clinical concept modifier recognition	• Recognition of **negations**, adaptation of NegEex • Recognition of expressions of clinical **certainty** • **Negated term** recognition (TENTES)
Other resources	• System of calculation **of similarity of medical texts** and / or to find similar cases (incl. text clustering): whole documents and sentence level • **Medical Machine translation** system (phrase-based and deep-learning based) • Medical Text annotation and **corpus labeling** system (Adaptation of AnnotateIt & BRAT) • Automatic **term mapping** & grounding using lexical similarity scores and ontology mapping • Annotation and **evaluation platform** for shared tasks (BeCalm – Markyt with Uni. Vigo)

“Big Data” en Biomedicina

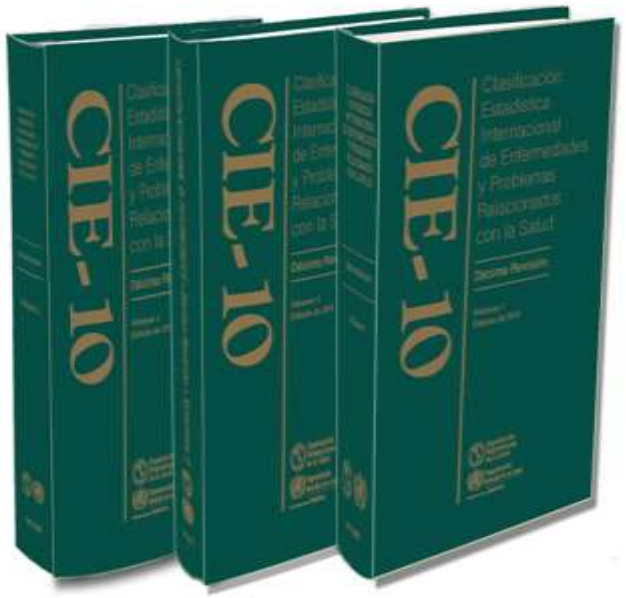
Interoperabilidad semántica y normalización de la historia clínica electrónica: modelo de producción de activos semánticos basados en estándares. Gonzalo Marco Cuenca (2016)

Capas del modelo de la red semántica UMLS

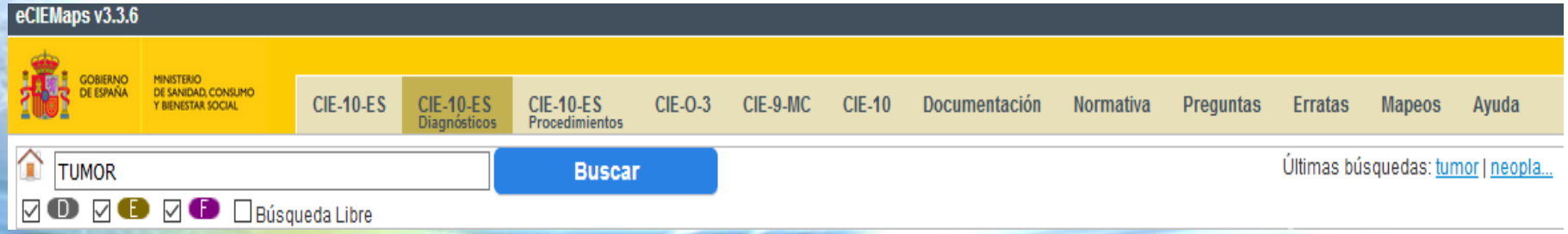
Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Taxonomía de enfermedades:

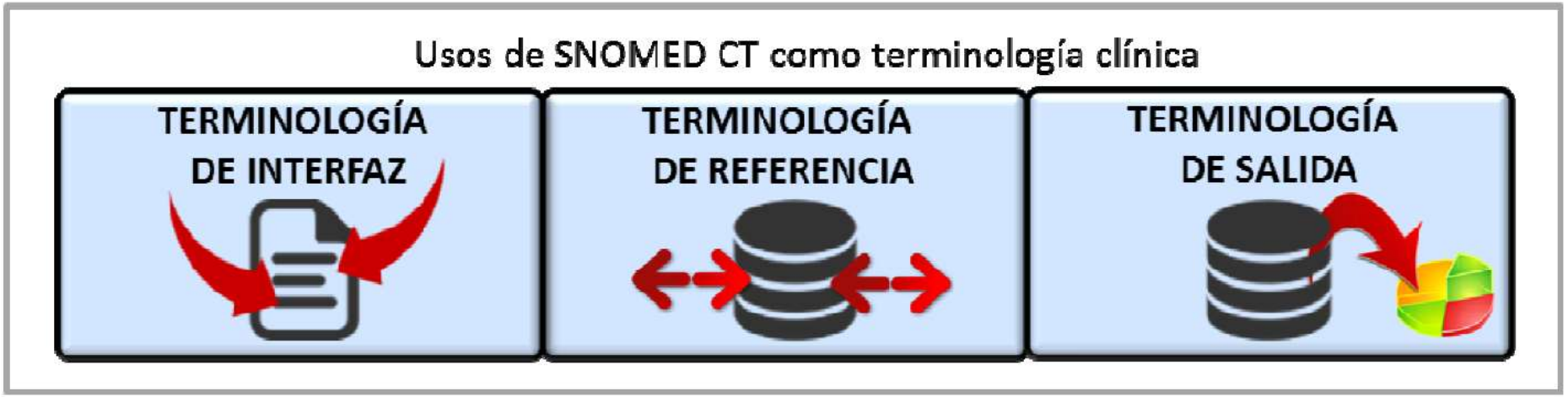
CIE-10



Capítulo	Códigos	Título
I	A00-B99	Ciertas enfermedades infecciosas y parasitarias
II	C00-D48	Neoplasias
III	D50-D89	Enfermedades de la sangre y de los órganos hematopoyéticos y otros trastornos que afectan el mecanismo de la inmunidad
IV	E00-E90	Enfermedades endocrinas, nutricionales y metabólicas
V	F00-F99	Trastornos mentales y del comportamiento
VI	G00-G99	Enfermedades del sistema nervioso
VII	H00-H59	Enfermedades del ojo y sus anexos
VIII	H60-H95	Enfermedades del oído y de la apófisis mastoides
IX	I00-I99	Enfermedades del sistema circulatorio
X	J00-J99	Enfermedades del sistema respiratorio
XI	K00-K93	Enfermedades del aparato digestivo
XII	L00-L99	Enfermedades de la piel y el tejido subcutáneo
XIII	M00-M99	Enfermedades del sistema osteomuscular y del tejido conectivo
XIV	N00-N99	Enfermedades del aparato genitourinario
XV	O00-O99	Embarazo, parto y puerperio
XVI	P00-P96	Ciertas afecciones originadas en el periodo perinatal
XVII	Q00-Q99	Malformaciones congénitas, deformidades y anomalías cromosómicas
XVIII	R00-R99	Síntomas, signos y hallazgos anormales clínicos y de laboratorio, no clasificados en otra parte
XIX	S00-T98	Traumatismos, envenenamientos y algunas otras consecuencias de causa externa
XX	V01-Y98	Causas extremas de morbilidad y de mortalidad
XXI	Z00-Z99	Factores que influyen en el estado de salud y contacto con los servicios de salud
XXII	U00-U99	Códigos para situaciones especiales



Diferentes usos de SNOMED CT como terminología clínica



Aspectos: terminología normativa - dominio clínico

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Interoperabilidad semántica y normalización de la historia clínica electrónica: modelo de producción de activos semánticos basados en estándares. Gonzalo Marco Cuenca (2016)

Criterios de selección de sistemas de codificación del proyecto epSOS.

Shared tasks and evaluation campaigns – Plan TL

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

	Tarea	Corpus	URL
MedTermMap Subtareas HPO Español e inglés	Tarea relacionada con mapeo automático de términos a conceptos en vocabularios controlados de la ontología HPO (Human Phenotype Ontology) versión en inglés y versión traducida a español.	• Guías de mapeo • Evaluación de calidad • Formato TSV y JSON • Expertos en terminología medica y fenotipos clínicos • EHR, Casos clínicos, terminologías internas	**TBD**
MedTermMap Subtarea SNOMED CT Español	Tarea relacionada con mapeo automático de términos a conceptos en vocabularios controlados de la ontología SNOMED CT versión en español.	• Guías de mapeo • Evaluación de calidad • Formato TSV y JSON • Expertos en terminología medica y fenotipos clínicos • EHR, Casos clínicos, terminologías internas	**TBD**

Generation of component- type resources: Automatic medical term recognition

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

CUTEXT - Cvalue Used To EXtract Terms

Introduction

The heavy use of medical terms has motivated the construction of large terminological resources for English, such as the Unified Medical Language System (UMLS) or the Open Biological and Biomedical Ontology (OBO) ontologies. Purely manual construction of terminological resources is by itself very valuable, but it constitutes a highly time-consuming process, it does not guarantee that included concepts or terms do actually align with the medical language and terms as they are being used in clinical documents by healthcare professionals and it requires constant update and revision due to changes and emergence of new biomedical concepts over time.

CUTEXT is a multilingual medical term extraction tool. It allows extracting terms in texts written in English, Spanish, Galician and Catalan.

The main characteristics of CUTEXT are the following:

- It is implemented in java, so it is multiplatform. It has been tested under Windows and Linux.
- It is multilingual: It has been tested in English, Spanish, Catalan and Galician, and it can be adapted easily to other languages by simply changing the lexical tag text file configuration.
- The entry documents can be in plain text or in pdf.
- It can be executed in graphic mode or by console (command line).
- It supports numerous configuration parameters, among the most important: the language, the tagger, the frequency and c-value thresholds and the entry of the document/s.
- The output is provided in plain text, in JSON format or/and in [BioC] (http://bioc.sourceforge.net/).

A more detailed description of the system can be found in the journal *Sociedad Española para el Procesamiento del Lenguaje Natural.*

https://github.com/PlanTL-SANIDAD/CUTEXT

Generation of terminological resources: medical entity recognition and machine translation

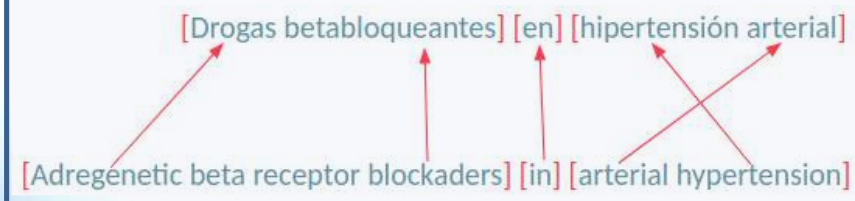
Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

1) *en* NER + *en-es* alignment

Use the **MeSpEN** to enrich **UMLS** automatically, that is to generate candidate term pairs in Spanish, both suggesting new terms or synonyms of already existing ones.

298,040 (en-es) **PubMed titles**

1. Identify **UMLS** terms in the English titles using **cTakes**.
2. Align the words of the titles in English to the words of the titles in Spanish (**Giza++**)
3. Using the previous alignment we detect the terms in the titles in English, and we assigned them to their corresponding **candidate** terms in Spanish.



Initial evaluation: a sample set of 200 candidate terms were manually validated by a domain expert.

- 47% were correct translations,
- 22% corresponded to either a more general term or more narrow term (hypernym/hyponym).
- the remaining pairs were either substrings of the correct translation or wrong translations. The average validation time per term was of just 2.03 seconds, using the MyMiner annotation tool.

2) Train NMT system + translate UMLS terms

Use a NMT system trained with medical corpus (Scielo *EN-ES* corpus) and use it to translate all the UMLS terms.

1. Use **DNorm** to get all disease mentions in the EN corpus.
2. Translate EN mentions into ES.

No evaluation yet (we are organising a manul evaluation)

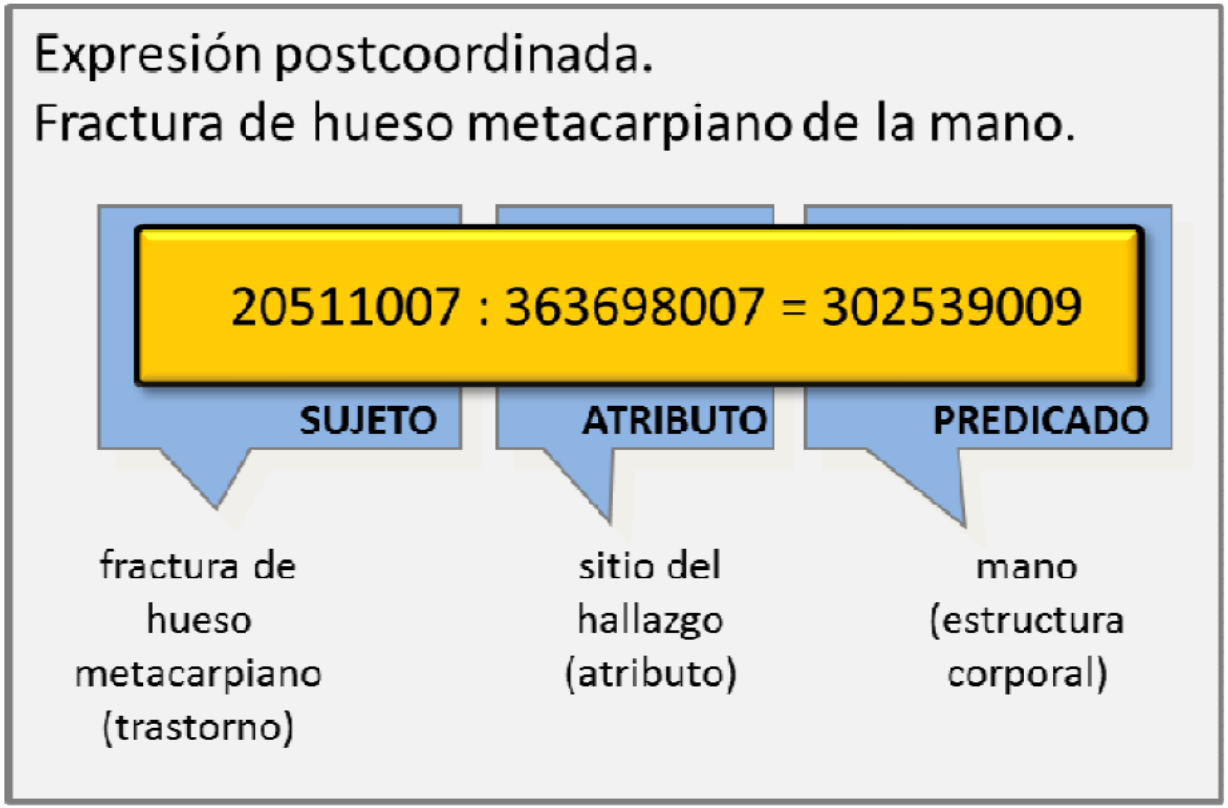
3) Train NMT system + translate corpus + NER + alignment

Train a NMT system with the Scielo EN-ES corpus

Spanish clinical cases corpus (extracted from 5,000 clinical cases articles)

1. Translate the Spanish clinical cases corpus into English.
2. Identify **UMLS** terms in English texts using **cTakes**.
3. Align the words of the texts in English to the words of the texts in Spanish (**Giza++**)
4. Using the alignment we detected the terms in the texts in English, and we assigned them to their corresponding **candidate** terms in Spanish.

Precoordinación y postcoordinación en SNOMED CT



Precoordinación y postcoordinación en SNOMED CT

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Expresión postcoordinada.
Fractura de hueso metacarpiano de la mano derecha.

Precoordinación y postcoordinación en SNOMED CT

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Principales características y diferencias entre SNOMED CT y las clasificaciones CIE y CIAP.	
SNOMED CT	**Clasificaciones (CIE/CIAP)**
Centrada en el paciente.	Centradas en la población.
Incluye descripciones para múltiples dominios de las Ciencias de la Salud: *Medicina, Farmacia, Enfermería, Laboratorio, etc.*, e incluye otros dominios de interés para la historia clínica electrónica.	Describen enfermedades, hallazgos, signos y síntomas y procedimientos del dominio de la Medicina.
Alta granularidad. Proporciona conceptos para describir situaciones clínicas con precisión.	Media o baja granularidad. Agrupan y comparan condiciones similares para interpretación estadística.
No contiene categorías residuales, impidiendo la existencia de conceptos ambiguos o indefinidos.	Contienen categorías residuales que facilitan la existencia de conceptos ambiguos e imprecisos (del tipo otros trastornos u otras enfermedades, como por ejemplo: otras enfermedades del aparato respiratorio).
Tienen un uso primario, siendo útiles para la toma de decisiones clínicas; sin despreciar otros posibles fines como la investigación, la estadística o la epidemiología.	Tienen un uso secundario, facilitando el análisis estadístico y epidemiológico.

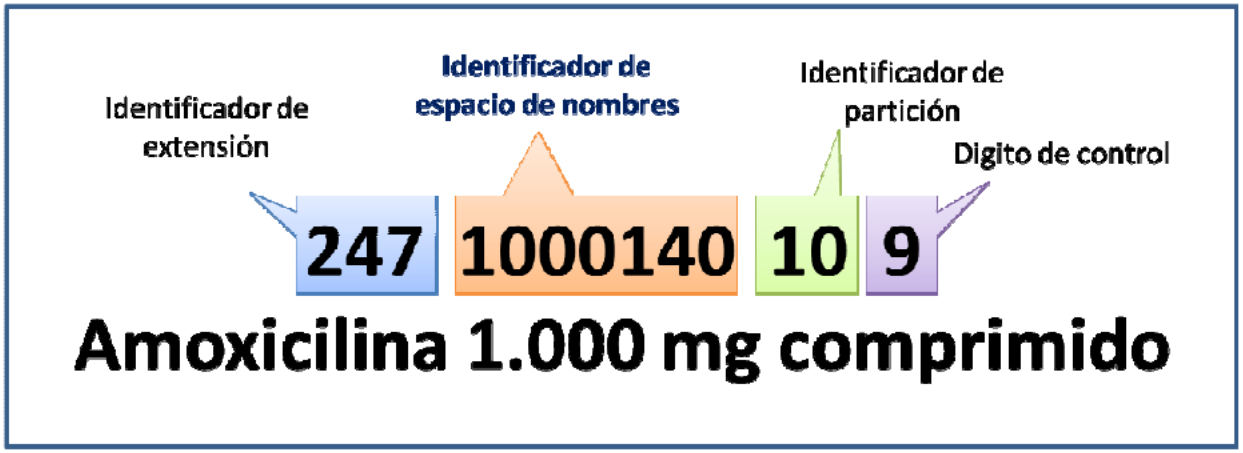
Interoperabilidad semántica y normalización de la historia clínica electrónica: modelo de producción de activos semánticos basados en estándares. Gonzalo Marco Cuenca (2016)

Precoordinación y postcoordinación en SNOMED CT

Principales características y diferencias entre SNOMED CT y las clasificaciones CIE y CIAP.	
SNOMED CT	**Clasificaciones (CIE/CIAP)**
Proporciona un lenguaje dinámico, flexible y adaptable por el usuario.	Proporcionan un lenguaje rígido, definido en el momento de su creación.
Incorpora un sistema de postcoordinación de términos para mejorar la precisión del significado.	No facilitan la postcoordinación.
Incluye referencias cruzadas o mapeos con las terminologías o clasificaciones más importantes en Ciencias de la Salud.	No incluyen referencias cruzadas.
Se incorporan cambios de forma rápida, gracias a su mecanismo de extensiones, en períodos cercanos (días, semanas y meses).	Se incorporan cambios de forma lenta y en períodos lejanos (por lo general años e, incluso, décadas).
Facilitan el intercambio de datos entre sistemas de información, siendo eficaces para entornos donde se requiere interoperabilidad semántica.	Pueden ser usadas por los sistemas de información para el intercambio de datos, pero no facilitan una interpretación precisa del significado.
Mantienen la trazabilidad de sus componentes a lo largo del tiempo y entre versiones, lo que permite consultar documentos históricos sin alteraciones.	No mantienen la trazabilidad entre versiones.

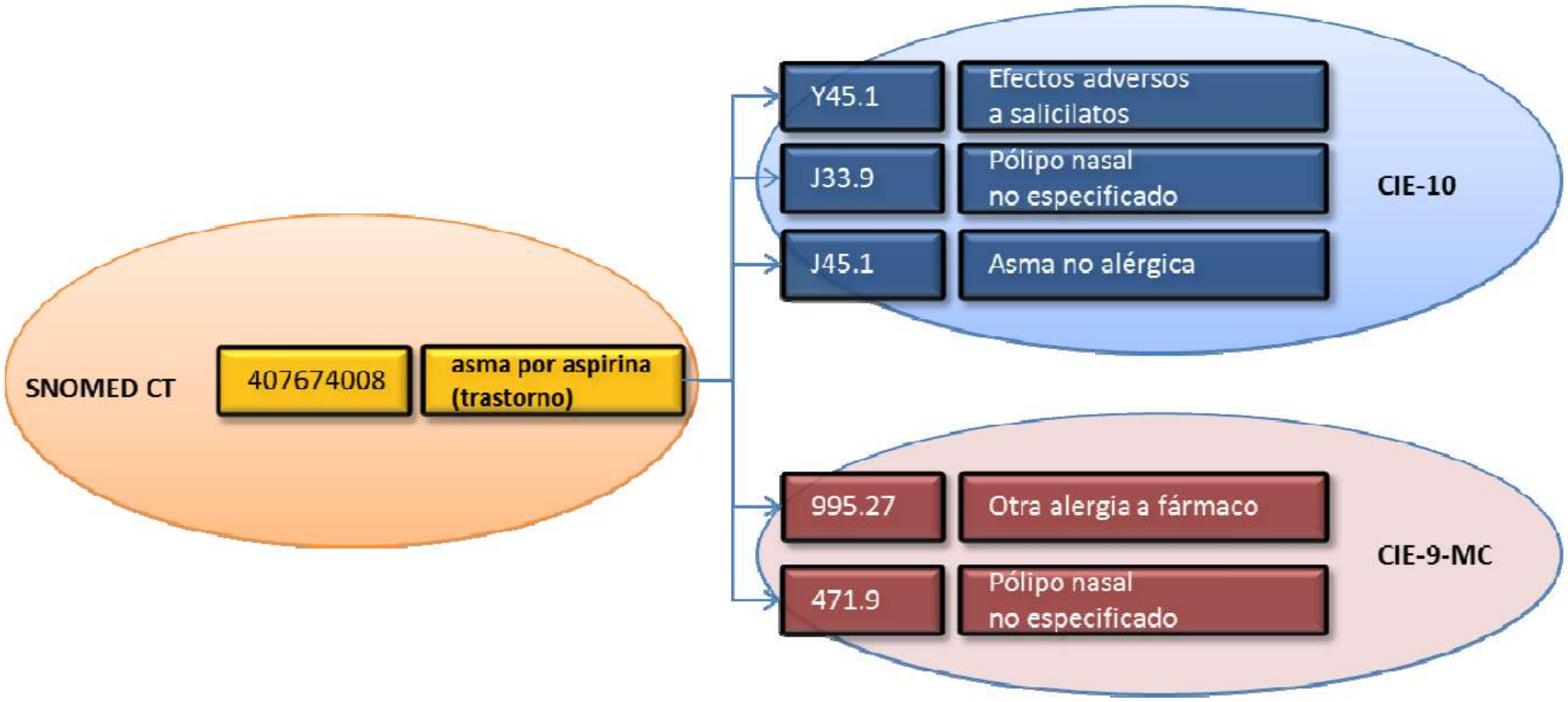
Interoperabilidad semántica y normalización de la historia clínica electrónica: modelo de producción de activos semánticos basados en estándares. Gonzalo Marco Cuenca (2016)

Identificador de concepto SNOMED CT



Componentes del identificador de concepto 2471000140109 | amoxicilina 1000 mg comprimido | que
forma parte de la extensión del Nomenclátor de prescripción de la AEMPS.

Identificador de concepto SNOMED CT



Establecimiento de referencias cruzadas 1 a n entre un concepto SNOMED CT y sus correspondencias con CIE-9-MC y CIE-10.

Principales formatos en vocabularios y terminologías

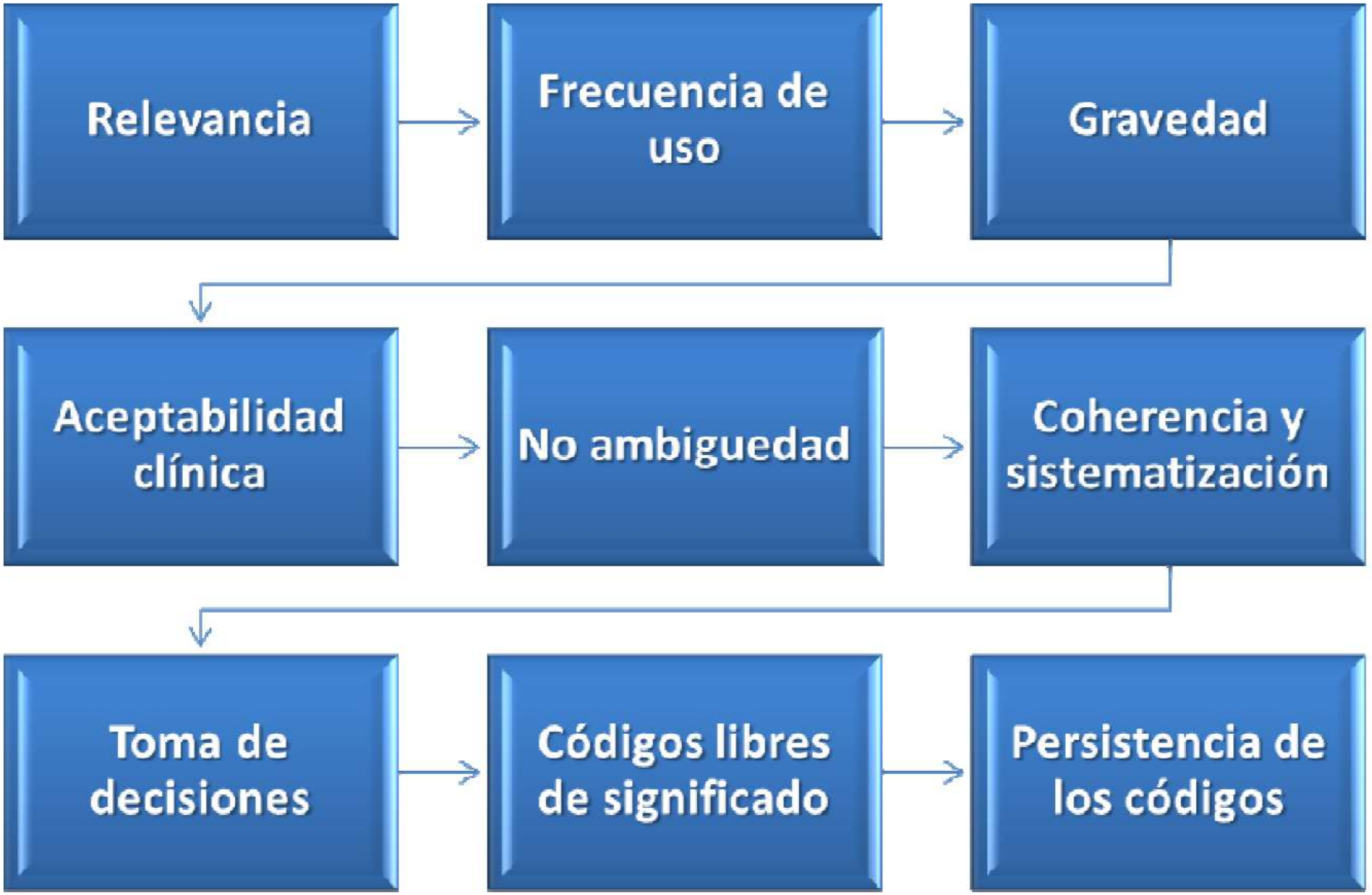
Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Formato	Tipo de recurso	Serialización
TMF/TBX	Terminologías	XML
LMF	Léxicos	LMF
SKOS	Tesauro	RDF
OWL	Ontologías	Varios
OBO	Ontologías	Propio
Ontolex	Léxicos basados en ontologías	RDF
TMX	Memorias de traducción	XML
XLIFF	Memorias de traducción	XML

Principales formatos y esquemas de anotación en PLN con especial énfasis en los usados en el dominio biomédico

Aspectos: terminología normativa - dominio clínico

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje



Interoperabilidad semántica y normalización de la historia clínica electrónica: modelo de producción de activos semánticos basados en estándares. Gonzalo Marco Cuenca (2016)

Criterios de selección de conceptos del proyecto epSOS

Tipos de recursos de terminología biomédica

- **Lista de términos**: conjunto de términos que representan de forma sintética entidades, no se representan relaciones
- **Taxonomía**: lista de términos organizados en un esquema jerárquico y por tanto, en categorías y subcategoría
- **Tesauro** – Taxonomía complementada con relaciones entre los términos
  - Relaciones de sinonimia o preferencia: entre el término preferido (TP) o descriptor y el término no preferido (TNP).
  - Relaciones jerárquicas de tipo partitivo (todo-parte) o clase-subclase; entre términos más amplios (TA) y más específicos (TE)
  - Relaciones asociativas: entre términos relacionados de forma pragmática, es decir, no de forma jerárquica ni de sinonímia
- **Ontologías**: pueden construirse a partir de Tesauros, codificados en un formato que pueda procesar un programa de software, en concreto un programa que pueda realizar inferencias. Para ello, debe estar enteramente basada en lógica formal

https://www.lluiscodina.com/taxonomias-ontologias/

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

www.disease-ontology.org

Proyecto que tiene como objetivo generar una estructura unificada para enlazar el conocimiento de las enfermedades humanas entre las distintas bases de datos:

Historia clínica
Secuenciación genoma
Microbioma
Dianas farmacológicas
Estudios de expresión génica
Modelos experimentales de enfermedad

Enlazar detecciones a UMLS

Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

- Mediante el conceptId de SNOMED podemos enlazar los términos a UMLS.

- Una vez enlazado a UMLS tenemos las entidades enlazadas a todo el contenido de UMLS:
  - *Systematized Nomenclature of Medicine – Clinical Terms* (SNOMED CT).
  - *Current Procedural Terminology* (CPT).
  - *The International Classification of Primary Care* (ICPC).
  - *Medical Dictionary for Regulatory Activities* (MedDRA).
  - *Medical Subject Headings* (MeSh).
  - *WHO Adverse Reactions Terminology* (WHOART).
  - ...