

Plan TL

Plan de Impulso de las
Tecnologías del Lenguaje



Corpus SPACCC de narrativa clínica: metodología y recursos

30
ANIVERSARIO
1989 - 2019

iic
instituto
de ingeniería
del conocimiento

Corpus SPACCC

Objetivo del proyecto

Poner a disposición de la comunidad científica un corpus biomédico exhaustivo que permita ejecutar tareas de Procesamiento de Lenguaje Natural (PLN) sobre grandes volúmenes de texto.

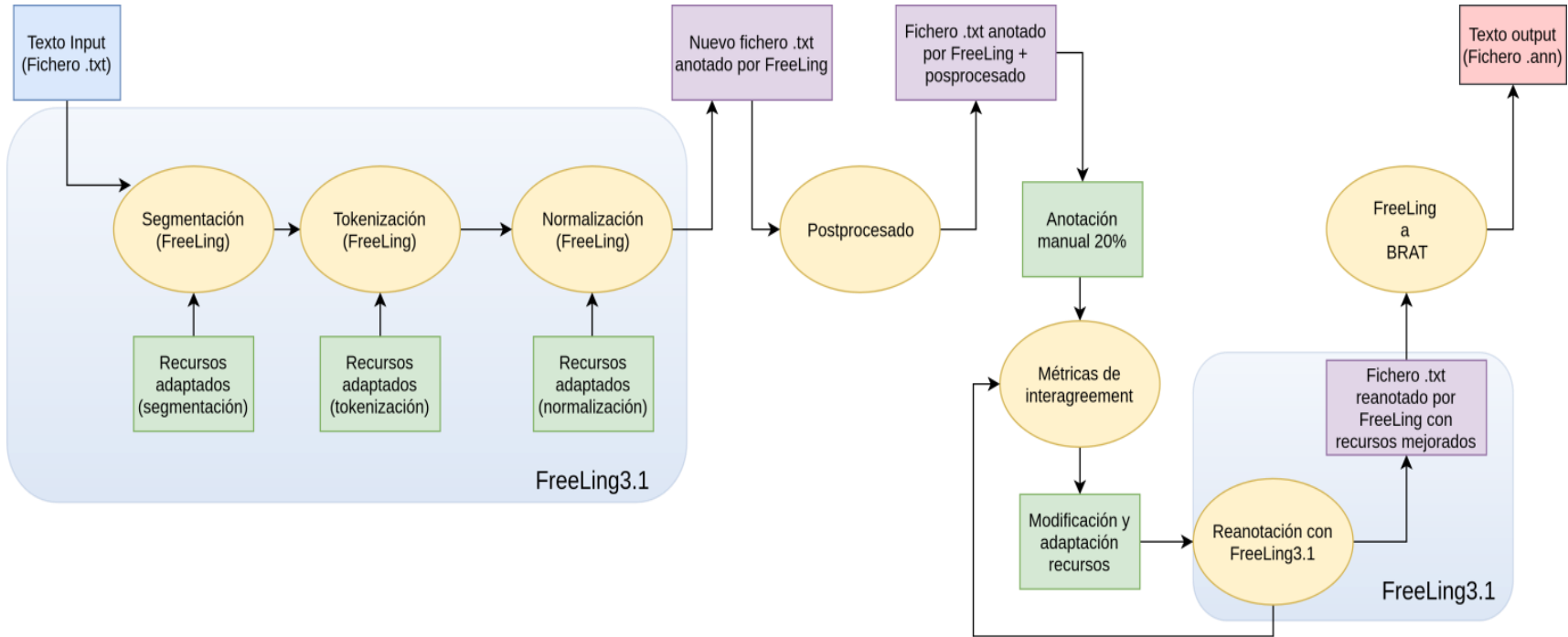
El corpus se compone de 1.000 **casos clínicos anonimizados** que se anotan en tres niveles lingüísticos:

- 1.- segmentación de oraciones
- 2.- segmentación de formas o *tokenización*
- 3.- etiquetado morfológico (POS).

Ejemplo de narrativa clínica

Varón de 36 años, sin antecedentes de interés, que fue estudiado en la consulta de medicina interna por presentar masa inguinoescrotal izquierda dolorosa a la palpación de dos meses de evolución, sin pérdida de peso ni síndrome miccional. A la exploración, los testes eran de tamaño y consistencia normales, con un cordón espermático izquierdo indurado y muy doloroso. La ecografía testicular fue normal. La CT de abdomen-pelvis reveló masa de 6 x 3 centímetros en el trayecto del cordón espermático izquierdo sin objetivarse imágenes de afectación retroperitoneal. Con el diagnóstico de tumor paratesticular izquierdo fue intervenido, encontrándose una masa en cordón espermático y realizándose biopsia intraoperatoria informada como proliferación neoplásica de aspecto miofibroblástico no linfomatosa, por lo que se realizó orquiectomía radical izquierda reglada. La anatomía patológica fue de rhabdomyosarcoma pleomórfico del cordón espermático, teste y epidídimo normales y negatividad de los márgenes de resección. Posteriormente el paciente ha recibido varios ciclos de quimioterapia con adriamicina e ifosfamida + MESNA. En las pruebas de imagen de control a los cuatro meses de la cirugía, no se objetivan recidivas tumorales.

Procedimiento proyecto completo



- Tres guías de anotación:
 - Split
 - Tokenización
 - POS
- Fijar los criterios de anotación e identificación de casos excepcionales.
 - Abreviaturas
 - Expresiones alfanuméricas
 - Nombres de medicamentos

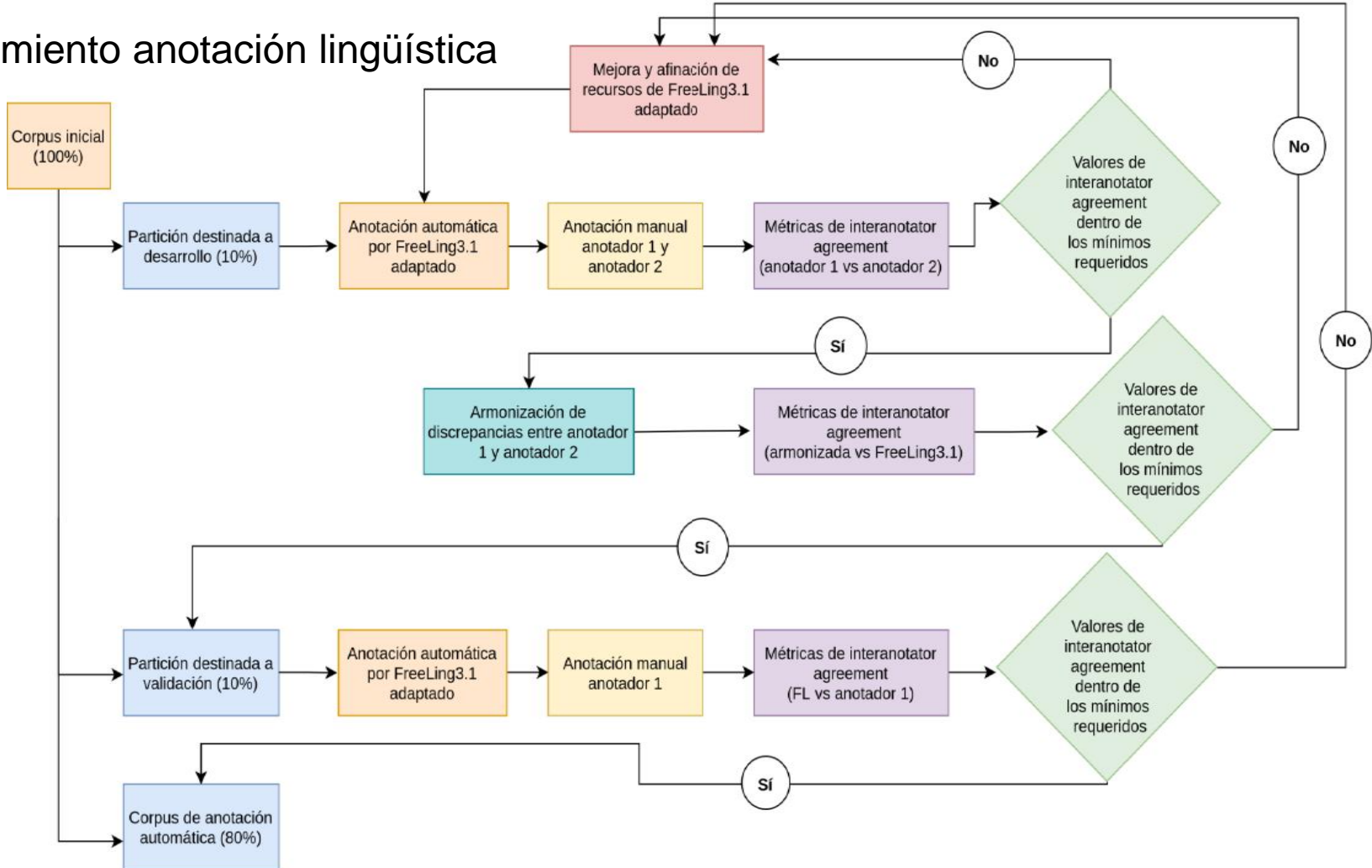
Guías de
anotación



Implementación de Freeling3.1

- Archivo SingleWords de dominio médico (16800 formas de 8998 formas no reconocidas por FreeLing)
- Tokenizer (243 siglas y abreviaturas)
- Usermap (557 expresiones alfanuméricas)

Procedimiento anotación lingüística



Anotación Manual

- Dos anotadores.
- 10% de desarrollo del corpus:
100 casos clínicos anotados sobre la anotación realizada con Freeling3.1 modificado para el dominio médico.
- Aplicación de criterios establecidos en las guías.
- Detección de posibles contraejemplos.



Valores mínimos de acuerdo entre anotadores requeridos:

Split	Token	POS
99%	98%	96%

Acuerdo entre anotadores



% de discrepancia entre anotadores y FreeLing3.1
(sobre 10% desarrollo):

	Split	Token	POS
A1 vs A2	99.79%	99.96%	98.84%
A1 vs FrL	99.51%	99.95%	98.13%
A2 vs FrL	99.72%	99.96%	98.53%

A1= Anotador 1
A2 = Anotador 2
FrL= FreeLing3.1

Acuerdo entre
anotadores II



- Resolución de discrepancias entre A1 y A2.
- Resultado: Gold Standard (GS)
- Cambios y mejoras en Freeling3.1

Armonización



- % acuerdo sobre 10% desarrollo (sin mejoras en FreeLing3.1).

	Split	Token	POS
GS vs FrL	99.51%	99.95%	97.89%

- % acuerdo sobre 10% desarrollo (con mejoras en FreeLing3.1).

	Split	Token	POS
GS vs FrL	99.51%	99.95%	98.1%
Requerido	99%	98%	96%

GS=Gold Standard
FrL= FreeLing3.1

Acuerdo entre
anotadores III



Anotación corpus de validación

- El 10% de validación es anotado por:
 - FreeLing3.1 adaptado al dominio médico (todas las mejoras incluidas).
 - Anotador 1 (siguiendo los criterios del GS)
- % de split, token y POS

	Split	Token	POS
GS vs FrL	99.37%	99.97%	98.85%
Requerido	99%	98%	96%

Corpus SPACCC

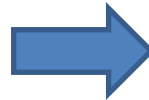
Metodología empleada

El éxito en estas métricas se debe al trabajo del equipo de expertos **lingüistas computacionales**, que ha estudiado a fondo casos específicos de la terminología para adaptar las herramientas de análisis del lenguaje estándar al dominio médico: en total, más de 300.000 palabras, 64.000 oraciones y 18.000 lemas diferentes se anotaron lingüísticamente, además de casos excepcionales del dominio médico (abreviaturas, unidades de medida, siglas, expresiones alfanuméricas...) para poner a disposición de la comunidad científica y la industria médica una rigurosa metodología de anotación.



Traducción a BRAT

Durante durante SPS00
el el DA0MS0
mismo mismo PI0MS000
se se P00CN000
detectó detectar VMIS3S0
una uno DIOFS0
tumoración tumoración NCFS000
de de SPS00
20 20 Z
mm milímetro NCMN000



T1	SP 0 7	Durante
#1	Norm T1	durante SPS00
T2	DA 8 10	el
#2	Norm T2	el DA0MS0
T3	PI 11 16	mismo
#3	Norm T3	mismo PI0MS000
T4	P0 17 19	se
#4	Norm T4	se P00CN000
T5	VM 20 27	detectó
#5	Norm T5	detectar VMIS3S0
T6	DI 28 31	una
#6	Norm T6	uno DIOFS0
T7	NC 32 42	tumoración
#7	Norm T7	tumoración NCFS000
A7	EM T7	
T8	SP 43 45	de
#8	Norm T8	de SPS00
T9	Z 46 48	20
#9	Norm T9	20 Z
T10	NC 49 51	mm
#10	Norm T10	milímetro NCMN000

Visualización en BRAT

The image shows a screenshot of the BRAT (Brat Rapid Annotation Tool) interface. The main text area contains the sentence: "1 Durante el mismo se detectó una tumoración bien delimitada e hipoecoica." Above the text, various words are annotated with colored boxes representing different parts of speech or entities. A tooltip window is open over the word "tumoración", displaying the following information:

- NC
- EM
- "tumoración"
- Norm: tumoración NCFS000
- ID: T7

Entregables

0. Corpus completo anotado con baselines.
1. FreeLing3.1 adaptado al dominio médico (en Docker).
2. Ficheros de FreeLing3.1 adaptados
3. Guías de anotación.
4. Corpus revisado en tres niveles por 2 expertos (20%).
5. Corpus completo anotado por FreeLing3.1 adaptado al dominio médico.
6. Informe de acuerdo entre anotadores.
7. Corpus completo anotado con BRAT.
8. Script de FreeLing3.1 a BRAT (en Docker).

Corpus SPACCC

Recursos accesibles:

<https://zenodo.org/communities/medicalnlp/search?page=1&size=20>

April 2, 2019 (v1.0.3) Software Open Access

View

SPACCC_POS-TAGGER

Felipe Soares; Aitor gonzalez-agirre;

[PlanTL/medicine/document annotation/NLP preprocessing/part-of-speech] Part-of-Speech Tagger for medical domain corpus in Spanish based on FreeLing.


Uploaded on April 2, 2019

3 more version(s) exist for this record

March 7, 2019 (v1.0.0) Software Open Access

View

SPACCC_Sentence-Splitter

Ander Intxaurreondo;  Martin;


[PlanTL/medicine/document/NLP preprocessing/sentence splitting] The sentence splitting model trained using the SPACCC_SPLIT corpus (https://github.com/PlanTL-SANIDAD/SPACCC_SPLIT). The model was trained using the 90% of the corpus (900 clinical cases) and tested against the 10% (100 clinical cases).

Uploaded on March 7, 2019

March 7, 2019 (v1.0.0) Software Open Access

View

SPACCC_Tokenizer

Ander Intxaurreondo;  Martin Krallinger;

[PlanTL/medicine/document/NLP preprocessing/tokenization] The tokenization model trained using the SPACCC_TOKEN corpus (https://github.com/PlanTL-SANIDAD/SPACCC_TOKEN). The model was trained using the 90% of the corpus (900 clinical cases) and tested against the 10% (100 clinical cases). This model

Uploaded on March 7, 2019

Gracias por su tiempo



www.iic.uam.es

C/ Francisco Tomás y Valiente, nº 11
EPS, Edificio B, 5ª planta
UAM Cantoblanco. 28049 Madrid
Tel.: (+34) 91 497 2323



Elementos gráficos de apoyo obtenidos en:

designed by  freepik.com

 pixabay.com

www.iic.uam.es

