GUÍA DE ANOTACIÓN DE TEXTOS MÉDICOS EN ESPAÑOL: TOKENIZACIÓN

Plan de impulso de las Tecnologías del Lenguaje

Nuria Aldama García[1]

Carmen Torrijos Caruda[1]

Montserrat Marimon[2]

Martin Krallinger[2,3]

[1]Instituto de ingeniería del conocimiento
[2]Centro Nacional de Supercomputación
[3]Centro nacional de Investigaciones Oncológicas

Julio 2018

Este estudio ha sido realizado dentro del ámbito del Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para el Avance Digital, que no comparte necesariamente los contenidos expresados en el mismo. Dichos contenidos son responsabilidad exclusiva de sus autores.

ÍNDICE

RESUMEN

Este documento presenta la herramienta utilizada para la tokenización de textos médicos en español, así como las convenciones que deben ser seguidas durante el proceso de anotación manual del corpus.

1 INTRODUCCIÓN

El Plan de Impulso de las Tecnologías del Lenguaje (Plan TL) tiene como objetivo fomentar el desarrollo del Procesamiento del Lenguaje Natural (PLN) y la Traducción Automática (TA) en lengua española y lenguas cooficiales. Para ello, el Plan TL define medidas que:

- Aumenten el número, calidad y disponibilidad de las infraestructuras lingüísticas en español y lenguas cooficiales.
- Impulsen la Industria del lenguaje fomentando la transferencia de conocimiento entre el sector investigador y la industria.
- Incorporen a la Administración como impulsor del sector de PLN.

Uno de los objetivos del proyecto es poner a disposición de la comunidad científica y la industria un corpus biomédico exhaustivo y con licencia abierta que permita ejecutar tareas de PLN sobre *big data* y replicar los experimentos. Este documento presenta la herramienta utilizada para la tokenización de textos médicos en español, así como las convenciones que deben ser seguidas durante el proceso de anotación manual del corpus. Consultar el documento *Metodología de anotación de textos biomédicos en español* para conocer los detalles relativos a los perfiles del autor de la guía y de los anotadores del corpus.

2 FREELING3.1

FreeLing [9] es una herramienta de análisis y etiquetado lingüístico que permite identificar el lenguaje al que pertenece una expresión lingüística, dividirla en oraciones, lematizarla y etiquetarla morfosintácticamente. Es una aplicación de código abierto para el procesamiento automático del lenguaje natural que proporciona una amplia gama de servicios de análisis lingüístico para una gran variedad de idiomas. Esta librería es personalizable y ampliable, y está fuertemente orientada al desarrollo de aplicaciones del mundo real en términos de velocidad y robustez. Además, permite al usuario analizar archivos de texto desde la línea de comandos. Por estos motivos, FreeLing es la herramienta que hemos elegido para la anotación de textos médicos en español, creando una

versión mejorada y adaptada al dominio médico a través de la modificación y el enriquecimiento de sus recursos de base.

2.1 REGLAS DE TOKENIZACIÓN POR DEFECTO (FREELING BASELINE)

El fichero de tokenización contiene dos secciones:

1. Reglas de tokenización:
    - iniciales
    - horas
    - expresiones alfanuméricas
    - URLs
    - emails
    - puntos suspensivos
    - comillas
    - palabras compuestas
    - abreviaturas
2. Abreviaturas declaradas: contiene una lista de 197 abreviaturas.

Estas reglas y abreviaturas están declaradas en el fichero tokenizer.dat

3 OTRAS HERRAMIENTAS ANALIZADAS

Las siguientes herramientas se citan en cada regla a lo largo de la guía de anotación, ya que se ha observado su tratamiento de los distintos problemas expuestos.

3.1 APACHE CTAKES+UIMA

CTAKES [13] es una herramienta de Procesamiento de Lenguaje Natural para la extracción de información a partir de registros clínicos electrónicos. Para su uso se utilizan pipelines personalizados, que consisten en modelos específicos entrenados con textos en inglés. En este caso hemos utilizado el *ClinicalPipeline* para observar el tratamiento de los distintos problemas de tokenización.

3.2 GENIA CORPUS

GENIA [8] es un corpus cuyo objetivo es desarrollar la extracción de información para el dominio específico de la biología molecular y las ciencias médicas. Está compuesto de títulos y abstracts de artículos académicos. Su función es el *mapping* entre las piezas de conocimiento y las estructuras lingüísticas. En este caso hemos utilizado las guías de anotación del corpus para observar el

tratamiento de los distintos problemas de tokenización. El proceso de anotación seguido para el corpus GENIA incluye los siguientes pasos [14]:

a. Los textos fueron tokenizados utilizando el tokenizador del Penn Treebank.

b. Preprocesado de textos mediante scrips en Perl centrados en la correcta tokenización de expresiones alfanuméricas propias del ámbito biomédico y asignación del POS para dichas expresiones.

c. Asignación del POS para el resto del corpus mediante una versión modificada del JunK tagger [6].

d. Proceso de corrección realizado por anotadores humanos.

4 REGLAS DE ANOTACIÓN MANUAL

Estas reglas proporcionan los detalles básicos de la anotación y las convenciones que deben ser seguidas durante el proceso de anotación manual del corpus. Las reglas se dividen en:

- Reglas generales: Reglas básicas que aplican a todos los procedimientos de tokenización.
- Reglas positivas: Reglas que aplican a casos específicos donde sí se anotan elementos conjuntamente como un solo token.
- Reglas negativas: Reglas que aplican a casos específicos donde los elementos se anotan como tokens separados. Se acompañan de ejemplos.
- Reglas ortográficas: Reglas que aplican a errores de tipografía. Se acompañan de ejemplos.

* En los ejemplos, la tokenización en unidades separadas se marca con el símbolo (|) cuando no se aporte una tabla explicativa.

4.1 REGLAS GENERALES (REGLAS-G)

- **G1. Abreviaturas con punto**

Las abreviaturas se tokenizan junto con el punto (.) que las acompaña, en el caso de que este aparezca ([4], [17]).

Ejemplo:

En el T.A.C. se evidencia un nódulo que infiltra en profundidad la grasa subcutánea y el músculo iliopsoas.

En la urografía intravenosa practicada se confirma el hallazgo ecográfico de defecto de replección en área lateral derecha, así como el estudio mediante T.A.C.

Forma	***Lema***	***POS***
R.T.U.	*r.t.u.*	*NCFS000*

CTAKES+UIMA ClinicalPipeline: Es inconsistente. Las abreviaturas con punto (.) que se encuentran a mitad de frase y que reconoce las segmenta como un solo token. Las abreviaturas con punto (.) que se encuentran a final de frase o que no reconoce, las segmenta como varios tokens.

GENIA (Annotation Guidelines, 2006): Los puntos de abreviaturas, siglas e iniciales se han eliminado del corpus inicial para evitar que constituyeran un SBS. Esto puede observarse mediante la descarga del corpus inicial en: http://www.geniaproject.org/genia-corpus/pos-annotation (descargando el fichero GENIAcorpus3.02p.tgz).

FreeLing3.1: Es consistente. Segmenta las abreviaturas con punto (.) como un solo token incluyendo los puntos dentro del mismo.

- **G2. Unidades de medida**

Las unidades de medida se tokenizan sin el punto (.) en aquellos casos en las que se vean acompañadas del mismo (ver regla O1) [17].

Ejemplo:

(…) y se administra de manera empírica, vancomicina 1g. junto a un antitérmico, el paciente permanece estable durante el mes siguiente.

Forma	***Lema***	***POS***
cm	*centímetro*	*NCMS000*
.	*.*	*Fp*

CTAKES+UIMA ClinicalPipeline: Es consistente. Separa el el punto (.) de la unidad de medida y lo tokeniza aparte.

GENIA (Annotation Guidelines, 2006): Los puntos de abreviaturas, siglas e iniciales se han eliminado del corpus inicial para evitar que constituyeran un SBS. Esto puede observarse mediante la descarga del corpus inicial en: http://www.geniaproject.org/genia-corpus/pos-annotation (descargando el fichero GENIAcorpus3.02p.tgz).

FreeLing3.1: Es consistente. Separa el el punto (.) de la unidad de medida y lo tokeniza aparte.

- **G3. Multiwords**

No se anotan multiwords en este proceso de anotación, ni de dominio general ni de dominio médico.

CTAKES+UIMA ClinicalPipeline: Realiza un etiquetado de entidades (enfermedades, síntomas, medicamentos, procedimientos médicos) desde una perspectiva semántica. No agrupa en chuncks.

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: Cuenta con un listado de multiwords que aplica por defecto que contiene expresiones de categoría cerrada como la locución preposicional 'en vez de' o expresiones de categoría abierta como 'fosa nasal' (ver regla I5).

4.2 REGLAS POSITIVAS (REGLAS-P)

- **P1. Tratamiento de abreviaturas y siglas**

Se consideran abreviaturas y siglas todas aquellas detectadas en el corpus y que aparezcan registradas en SNOMED, y cada una constituye un solo token.

Ejemplo:

Forma	***Lema***	***POS***
IMC	*IMC*	*NCMS000*

CTAKES+UIMA ClinicalPipeline: contiene distintos listados de siglas y abreviaturas en inglés pertenecientes a diversos dominios como por ejemplo PersonTitleAnnotator.xml o NamesandGovernmentOfficials_TAE.xml

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: las siglas y abreviaturas se incluyen en el fichero de tokenización y se normalizan en el fichero de normalización. De este modo se tokenizan y normalizan correctamente.

- **P2. Unidades de medida complejas separadas por barras: *slash* (/) o *backslash* (\)**

Las unidades de medida complejas separadas por barras (/, \) se separan en tokens diferentes [17].

Ejemplo:

Forma	***Lema***	***POS***
mg	*miligramo*	*NCMS000*
/	*/*	*Fh*
ml	*mililitro*	*NCMS000*

CTAKES+UIMA ClinicalPipeline: segmenta las unidades de medida complejas separadas por barras (/, \) en tantos tokens como elementos la compongan.

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: segmenta las unidades de medida complejas separadas por barras (/, \) en tantos tokens como elementos la compongan.

- **P3. Números ordinales**

Los números ordinales junto con los símbolos 'ª' y 'º' se tratan como un solo token. También se consideran un solo token cuando aparezcan con el formato '3.a' o '3a'.

Ejemplo: 3ª

Forma	***Lema***	***POS***

3ª	*3ª*	*Z*
3a	*3a*	*Z*
3.a	*3.a*	*Z*

CTAKES+UIMA ClinicalPipeline: Los números ordinales abreviados '3º' o '3o' se segmentan como un solo token. Los números ordinales abreviados '3.o' son segmentados en dos tokens (3./o)

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: Cualquiera de los tres formatos de representación de números ordinales incluidos en la tabla son tokenizados como una única unidad.

- **P4. Fechas en formato numérico**

Se tokenizan de forma conjunta las fechas en formato numérico, con guiones (-) o barras inclinadas (/, \) [17].

Ejemplo: 12-5-2005

Forma	***Lema***	***POS***
12-5-2005	*12-5-2005*	*Z*
12/5/2005	*12/5/2005*	*Z*

CTAKES+UIMA ClinicalPipeline: segmenta los formatos numéricos de fecha en tantos tokens como elementos la compongan.

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: segmenta los formatos numéricos de fecha como un solo token.

- **P5. Hora en formato numérico**

Se tokenizan de forma conjunta las horas en formato numérico, con dos puntos (:) o un punto (.).

Forma	*Lema*	*POS*
20:35	*20:35*	*Z*
19.42	*19.42*	*Z*

CTAKES+UIMA ClinicalPipeline: Segmenta los formatos de hora separados por dos puntos (:) en tres tokens (20 | : | 35). Segmenta los formatos de hora separados por un punto (.) en un solo token.

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: Segmenta tanto los formatos de hora separados por dos puntos (:) como los formatos de hora separados por un punto (.) en un solo token.

- **P6. Palabras unidas por guion**

Las palabras unidas por un guion (-) sin espacios son reconocidas como un solo token ([15], [16]).

Ejemplo: Budd-Chiari

Forma	*Lema*	*POS*
Budd-Chiari	*Budd-Chiari*	*NP00000*

CTAKES+UIMA ClinicalPipeline: Las palabras unidas por un guion (-) sin espacios, son segmentadas en tantos tokens como elementos la compongan (Budd | - | Chiari).

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: Las palabras unidas por un guion (-) sin espacios, son segmentadas como un único token.

- **P7. Palabras separadas por barras: *slash* (/) y *back slash* (\)**

Por defecto las palabras, siglas, abreviaturas o unidades de medida separadas por las barras (/, \) se separan en tokens diferentes.

Ejemplos:

Forma	***Lema***	***POS***
cigarrillos	*cigarrillos*	*NCMP000*
/	*/*	*Fh*
día	*día*	*NCMS000*

CTAKES+UIMA ClinicalPipeline: Las palabras separadas por barras son tokenizadas en el número de elementos que compongan la expresión (cigarrillos | / | día).

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: Las palabras separadas por barras son tokenizadas en el número de elementos que compongan la expresión (cigarrillos | / | día).

- **P8. Complejos numéricos separados por barras: *slash* (/) y *back slash* (\)**

Los complejos numéricos separados por barras (/, \) se tokenizan como una única unidad.

Ejemplos:

Forma	***Lema***	***POS***
140/100	*140/100*	*Z*

CTAKES+UIMA ClinicalPipeline: Los complejos numéricos separados como una barra (/, \) son segmentados en tantas unidades como elementos la compongan ('140 | / | 100').

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: Los complejos numéricos separados como una barra (/, \) son segmentados en un solo token por la regla de tokenización NUMBERS (ver regla I4).

- **P9. Paréntesis, llaves, comillas, exclamaciones e interrogaciones.**

Los paréntesis (() ()), llaves ({) (}), comillas (“) (”), exclamaciones (¡) (!) e interrogaciones (¿) (?), y en general todos los signos de puntuación de apertura y cierre se tokenizan siempre como tokens separados [16].

Ejemplo: (ECR-NMDA)

Forma	***Lema***	***POS***
(	*(*	*Fpa*
ECR-NMDA	*ECR-NMDA*	*NP00000*
)	*)*	*Fpt*

CTAKES+UIMA ClinicalPipeline: Los paréntesis (() ()), llaves ({) (}), comillas (“) (”), exclamaciones (¡) (!) e interrogaciones (¿) (?) se tokenizan como tokens separados.

GENIA (Annotation Guidelines, 2006): Los paréntesis (() ()) y los dos puntos (:) son tokenizados como elementos individuales. No especifica para el resto de signos de puntuación.

FreeLing3.1: Los paréntesis (() ()), llaves ({) (}), comillas (“) (”), exclamaciones (¡) (!) e interrogaciones (¿) (?) se tokenizan siempre como tokens separados.

- **P10. Emails y URLs**

Las direcciones de email y las URLs se consideran un solo token ([5], [17]).

Ejemplo: http://nefrochus.villaweb.es/en/

Forma	***Lema***	***POS***
http://nefrochus.villaweb.com/es/en/	*http://nefrochus.villaweb.es/en/*	*NP00000*

CTAKES+UIMA ClinicalPipeline: Segmenta las URLs en 5 tokens (‘http://nefrochus.villaweb.com | / | es | / | en’)

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: Segmenta las URLs en 5 tokens (‘http://nefrochus.villaweb.com | / | es | / | en’) (ver regla I6).

- **P11. Afijos en casos de coordinación**

Los afijos que aparezcan como palabras separadas en casos de coordinación se tratan en un solo token [17]

Ejemplo: pre y postoperatorio

Forma	***Lema***	***POS***
pre	*pre*	*RG*

CTAKES+UIMA ClinicalPipeline: Segmenta los afijos como palabras separadas en casos de coordinación.

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: Segmenta los afijos como palabras separadas en casos de coordinación.

- **P12. Afijos unidos a la palabra**

Los afijos que aparezcan unidos a la palabra se tokenizan junto con la misma.

Ejemplo: postoperatorio

Forma	***Lema***	***POS***
postoperatorio	*postoperatorio*	*AQ0MS0*

CTAKES+UIMA ClinicalPipeline: Segmenta los afijos unidos a la palabra que modifican como un solo token.

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: Segmenta los afijos unidos a la palabra que modifican como un solo token.

- **P13. Cadenas alfanuméricas**

Las cadenas alfanuméricas correspondientes a entidades médicas como cromosomas o identificadores de bases de datos se tratan como tokens separados ([3], [5], [15], [16]).

Ejemplo: CD4/CD8

Forma	*Lema*	*POS*
CD4	*CD4*	*NC00000*
/	*/*	*Fh*
CD8	*CD8*	*NC00000*

CTAKES+UIMA ClinicalPipeline: Segmenta las cadenas alfa numéricas en tantos elementos como la compongan.

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: Segmenta las cadenas alfa numéricas en un solo token (ver regla I4).

- **P14. Cantidades y dosis**

El número y la unidad de medida se separan siempre en dos tokens diferentes.

Ejemplo: 70mg

Forma	*Lema*	*POS*
70	*70*	*Z*
mg	*miligramo*	*NCMS000*

CTAKES+UIMA ClinicalPipeline: Separa el número y la unidad de medida tanto si éstas están separadas por un espacio (’70 mg’) como si se encuentran juntas (‘70mg’).

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: No separa el número de la unidad de medida si la cifra y la unidad de medida no están separadas por un espacio (‘70mg’). Separa el número y la unidad de medida si éstas están separadas por un espacio (’70 mg’). Cuando aparezcan unidos en el corpus se separan por la regla NUMBERS declarada en el fichero de reglas de tokenización (ver regla I1).

- **P15. Nombres propios**

En el caso de que vayan escritos en mayúscula inicial, tanto en el caso de nombre + apellido como en el caso de inicial + apellido o apellido + inicial, los nombres propios se tratan como un solo token.

Ejemplo:

Forma	***Lema***	***POS***
T. Millin	*T_._Millin*	*NP00000*
Schindler L.	*Schindler_L_.*	*NP00000*
Bristol-Myers Squibb	*Bristol_-_Myers_Squibb*	*NP00000*

CTAKES+UIMA ClinicalPipeline: Es inconsistente. Tokeniza las secuencias 'inicial + punto (.) + apellido' y 'nombre + apellido' en dos unidades ('T. | Millin' 'Bristol-Myers | Squibb'). Tokeniza la secuencia 'nombre propio + inicial + punto (.)' como tres elementos independientes ('Schindler | L | . ').

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: Tokeniza cada uno de los formatos de nombres propios incluidos en la tabla superior como un único token.

- **P16. Tratamiento de isótopos**

Las expresiones del tipo Ca(2+) se anotan en un solo token.

Ejemplo:

Forma	***Lema***	***POS***
Ca(2+)	*Ca(2+)*	*NP00000*

CTAKES+UIMA ClinicalPipeline: Tokeniza los paréntesis como elementos independientes.

GENIA (Annotation Guidelines, 2006): Tokeniza los paréntesis como elementos independientes.

FreeLing3.1: Tokeniza los paréntesis como elementos independientes (ver regla I7).

4.3 REGLAS NEGATIVAS (REGLAS-N)

- **N1. Palabras separadas por espacio guion espacio.**

Las palabras unidas por un guion (-) pero separadas por espacios no se tokenizan como un solo elemento sino como varios tokens diferenciados [17].

Ejemplo: ansiolítico – antidepresivo

Forma	***Lema***	***POS***
ansiolítico	*ansiolítico*	*NCMS000*
-	-	*Fg*
antidepresivo	*antidepresivo*	*AQ0MS0*

CTAKES+UIMA ClinicalPipeline: Segmenta la expresión en tres tokens diferenciados (‘ansiolítico | – | antidepresivo’).

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: Segmenta la expresión en tres tokens diferenciados (‘ansiolítico | – | antidepresivo’).

- **N2. Cantidades y dosis**

El número y la unidad de medida se separan siempre en dos tokens diferentes.

Ejemplo: 70mg

Forma	***Lema***	***POS***
70	*70*	*Z*
mg	*miligramo*	*NCMS000*

CTAKES+UIMA ClinicalPipeline: Separa el número y la unidad de medida tanto si éstas están separadas por un espacio (’70 mg’) como si se encuentran juntas (‘70mg’).

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: No separa el número de la unidad de medida si la cifra y la unidad de medida no están separadas por un espacio ('70mg'). Separa el número y la unidad de medida si éstas están separadas por un espacio ('70 mg'). Cuando aparezcan unidos en el corpus se separan por la regla NUMBERS declarada en el fichero de reglas de tokenización (ver regla I1).

- **N3. Dos puntos**

Las palabras, números o cadenas alfanuméricas que vayan unidas por dos puntos con espacio o sin espacio se tokenizarán por separado:

Ejemplo: estadio:t4m2n0

Forma	***Lema***	***POS***
estadio	*estadio*	*NCMS000*
:	*:*	*Fd*
t4m2n0	*t4m2n0*	*NC00000*

CTAKES+UIMA ClinicalPipeline: Tokeniza por separado los dos puntos cuando se encuentran en medio de palabras, números o cadenas alfanuméricas.

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: Tokeniza por separado los dos puntos cuando se encuentran en medio de palabras, números o cadenas alfanuméricas. Para evitar en la anotación automática que esta regla separe el formato horario hh:mm o bien hh:mm:ss se ha añadido la regla HOURS en el fichero de reglas de tokenización (ver regla de implementación I8).

- **N4. Asterisco (*)**

Las palabras, números o cadenas alfanuméricas que vayan unidas por un asterisco con espacio o sin espacio se tokenizarán por separado:

Ejemplo: p.gly1175valfs*64

Forma	***Lema***	***POS***

p.gly1175valfs	*p.gly1175valfs*	*N000000*
*	*	*Fz*
64	*64*	*Z*

CTAKES+UIMA ClinicalPipeline: Tokeniza por separado los asteriscos cuando se encuentran en medio de palabras, números o cadenas alfanuméricas.

GENIA (Annotation Guidelines, 2006): Tokeniza de manera conjunta las cadenas alfanuméricas cuando encuentra un asterisco en medio.

FreeLing3.1: Tokeniza por separado los asteriscos cuando se encuentran en medio de palabras, números o cadenas alfanuméricas.

- **N5. Signo igual (=)**

Las palabras, números o cadenas alfanuméricas que vayan unidas por un signo igual con espacio o sin espacio se tokenizarán por separado:

Ejemplo: EAV=8-9

Forma	***Lema***	***POS***
EAV	*EAV*	*NC00000*
=	*=*	*Fz*
8-9	*8-9*	*Z*

CTAKES+UIMA ClinicalPipeline: Tokeniza por separado los signos de igual (=) cuando se encuentran en medio de palabras, números o cadenas alfanuméricas.

GENIA (Annotation Guidelines, 2006): No especifica.

FreeLing3.1: Tokeniza por separado los signos de igual (=) cuando se encuentran en medio de palabras, números o cadenas alfanuméricas.

4.4 REGLAS ORTOGRÁFICAS (REGLAS-O)

- **O1. Unidades de medida**

Las unidades de medida se escriben sin punto, por lo que el punto se tokeniza siempre por separado, nunca como parte de la abreviatura.

Forma	***Lema***	***POS***
mg	*miligramo*	*NCMS000*
.	*.*	*Fh*

4.5 IMPLEMENTACIÓN DE LAS REGLAS EN ANOTACIÓN AUTOMÁTICA (REGLAS-I)

Estas reglas proporcionan los detalles de implementación de las reglas de anotación manual en el proceso de anotación automática de FreeLing3.1.

- **I1. Unidades de medida**

Se incorpora una regla de tokenización que permite el mejor reconocimiento de los números para una tokenización separada de las unidades de medida. En el caso de “5,5mg” la versión de base de FreeLing lo interpreta como una expresión alfanumérica en un solo token, mientras que con esta regla tendremos dos token separados: "5,5" y "mg", al haberse reconocido el primero como un número. La regla se declara como **NUMBERS** en el fichero tokenizer.dat y consiste en la siguiente expresión regular:

[+-]?[0-9]+({SYMNUM}[0-9]+)*{SYMNUM}?

- **I2. Expresiones de temperatura**

Se incorpora una regla de tokenización que permite el mejor reconocimiento de las expresiones de temperatura (ºC), mediante la tokenización separada del numeral y la unidad de medida. En el caso de “36ºC” la versión de base de FreeLing lo interpreta como una expresión alfanumérica en un solo token, mientras que con esta regla tendremos dos token separados: “36” y “ºC”, al haberse reconocido el segundo como la unidad de medida de los grados centígrados. La regla se declara como **DEGREES** en el fichero tokenizer.dat y consiste en la siguiente expresión regular:

([+-]?[0-9]+(?:{SYMNUM}[0-9]+)*)(º{SPACE}?C)

- **I3. Abreviaturas**

A partir de una lista proporcionada por el CNIO de 6.139 abreviaturas lematizadas, se ha realizado una revisión para eliminar repetidos, controlar ambigüedades y comprobar la normalización con la base de datos de SNOMED. Tras esta revisión el conjunto se ha reducido a un total de 755 abreviaturas y siglas que se declaran en el fichero de tokenización de FreeLing tokenizer.dat.

- **I4. Complejos numéricos separados por barras**

Los complejos numéricos separados por barras se anotan en un solo token por la regla de tokenización NUMBERS. La regla se declara como **NUMBERS** en el fichero tokenizer.dat y consiste en la siguiente expresión regular:

[+-]?[0-9]+({SYMNUM}[0-9]+)*{SYMNUM}?

- **I5. Fichero de multiwords FreeLing3.1**

Se elimina el fichero de multiwords básico de FreeLing3.1 para evitar que las multiwords que este contiene aparezcan tokenizadas conjuntamente en la anotación automática del texto.

- **I6. Tratamiento de URLs**

Los slash (/) incluidos en las URLs se tokenizan junto con la URL mediante la modificación de la regla de tokenización de FreeLing para URLs en el fichero tokenizer.dat, que una vez modificada consiste en la siguiente expresión regular:

((mailto:|(news|http|https|ftp|ftps)://)[\w\.\-]+|^(www(\.[\w\-]+)+))[/\w\.\-]*

- **I7. Tratamiento de isótopos**

Se incorpora una regla de tokenización que permite anotar de manera automática como un solo token las expresiones del tipo Ca(2+). La regla se declara como **ELEMENTS** en el fichero tokenizer.dat y consiste en la siguiente expresión regular:

[A-Z][a-z]?\([0-9]+[\+\-]\)

- **I8. Tratamiento de horas**

Se incorpora una regla de tokenización que permite anotar de manera automática como un solo token las expresiones horarias del tipo hh:mm o bien hh:mm:ss. La regla se declara como **HOURS** en el fichero tokenizer.dat y consiste en la siguiente expresión regular:

([0-9][0-9]?:[0-9][0-9](:[0-9][0-9])?)[^0-9]

5 BIBLIOGRAFÍA

[1] Barret, N. & Weber-Jahnke, J. (2014) A token centric part-of-speech tagger for biomedical text. *Artificial Intelligence in Medicine*, 61, 11-20.

[2] Campillos, L., Deléger, L., Grouin, C., Hamoon, T., Ligozat, A.L. & Névéol, A. (2018) A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annotated Text corpus (MERLOT). *Language Resources & Evaluation, 52:571-601*.

[3] Fan, J.W., Yang, E.W., Jiang, M., Prasad, R., Loomis, R.M., Zisook, D.S., Denny, J.C., Xu, H. & Huang, Y. (2013) Syntactic parsing of clinical text: guideline and corpus development with handling ill-formed sentences. *Journal of the American Medical Informatics Association: JAMIA*, *20*(6), 1168–1177.

[4] Griffis, D., Shivade, C., Fosler-Lussier, E. & Lai, A.M. (2016) A Quantitative Evaluation of Sentence Boundary Detection for the Clinical Domain. *AMIA Summits on Translational Science Proceedings*, 88–97.

[5] He, Ying & Kayaalp, M. (2006) A comparison of 13 Tokenizers on MEDLINE. Technical Report. Available at: https://lhncbc.nlm.nih.gov/publication/lhncbc-tr-2006-003 Access date: 8/06/2018

[6] Kazama, J., Miyao, Y. & Tsujii. J. (2001) A Maximum Entropy Tagger with Unsupervised Hidden Markov Models. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*. November 2001. Tokyo, Japan. 333--340.

[7] Kim, J., Ohta, T. Teteisi, Y. & Tsujii, J. (2006) Genia Corpus Manual: Encoding schemes for the corpus and annotation. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.106.9947&rep=rep1&type=pdf Access date: 21/06/2018.

[8] Kim J.D., Ohta T., Tateishi Y., & Tsujii J. (2003) GENIA corpus - a semantically annotated corpus

for bio-textmining. *Bioinformatics*, 19:suppl. 1):180–i182.

[9] Padró, L & Stanilovsky, E. (2012) FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*. May 2012. Istanbul, Turkey. http://nlp.lsi.upc.edu/publications/papers/padro12.pdf Accessed date: 05/07/2018

[10] Pakhomov, S.V., Coden, A. & Chute, C.G. (2006) Developing a corpus of clinical notes manually annotated for part-of-speech. *International Journal of Medical Informatics*, 75, 418-429.

[11] RAE (2005) Signos ortográficos. *Diccionario panhispánico de dudas Real Academia Española http://lema.rae.es/dpd/srv/search?id=qXGSxldBKD6hqrTMMo*

[12] Sætre, R.,Yoshida, K., Yakushiji, A., Miyao, Y., Matsubayashi, Y. & Ohta, T. AKANE System: Protein-Protein Interaction Pairs in BioCreAtIvE2 Challenge, PPI-IPS subtask. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*. April 2007. 209--212.

[13] Savova, G. K., Masanz, J. J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K. C. & Chute, C. G. (2010) Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, *17*(5), 507–513. http://doi.org/10.1136/jamia.2009.001560

[14] Teteisi, Y. & Tsujii, J. (2006) Genia Annotation Guidelines for Tokenization and POS Tagging. Available at: http://www.nactem.ac.uk/tsujii/papers/yucca/GENIA_Guidelines_POS.pdf.4 Access date: 21/06/2018

[15] Teteisi, Y. & Tsuji, J. (2004) Part-of-Speech Annotation of Biology Research Abstracts. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004) May 2004, Lisbon, Portugal* 1267-1270. http://www.nactem.ac.uk/aigaion2/index.php?/publications/show/129 Accessed date: 02/07/2018.

[16] Tomanek, K., Wermter, J. & Hahn, U. (2007) Sentence and Token Splitting On Conditional Random Fields. Available at: https://pdfs.semanticscholar.org/5651/b25a78ac8fd5dd65f9c877c67897f58cf817.pdf Access date: 9/06/2018

[17] Warner, C., Lanfranchi, A., O'Gorman, T., Howard, A., Gould, K. & Regan, M. (2012) Bracketing Biomedical Text: An Addendum to Penn Treebank II Guidelines. Available at: https://clear.colorado.edu/compsem/documents/treebank_guidelines.pdf Access date: 11/06/2018

6 GLOSARIO DE SIGLAS Y ACRÓNIMOS

Plan TL	Plan de Impulso de las Tecnologías del Lenguaje
PLN	Procesamiento del Lenguaje Natural
TA	Traducción Automática