GUÍA DE ANOTACIÓN DE TEXTOS MÉDICOS EN ESPAÑOL:

SEGMENTACIÓN DE FRASES

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

ÍNDICE DE TABLAS

RESUMEN

Este documento presenta la herramienta utilizada para la segmentación de frases 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 segmentación de frases 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 SEGMENTACIÓN DE FRASES (FREELING3.1 BASELINE)

Para realizar el *sentence splitting*, FreeLing3.1 contiene una serie de reglas básicas que se detallan a continuación y que establecen el *baseline*:

1. Por defecto no se permite la segmentación de frases en texto comprendido entre dos marcadores. Los marcadores declarados son: " "( ){ }/* */. Esta opción se puede deshabilitar.

2. Por defecto no hay ningún valor asignado al máximo de palabras que se permiten antes de realizar una segmentación de frase dentro de los marcadores. Este valor se puede modificar.

3. La interrogación de cierre (?) constituye un finalizador de frase ambiguo, por lo que no siempre supone un final de frase, sino que únicamente segmenta si la palabra que le sigue comienza por mayúscula.

4. El punto (.), la admiración de cierre (!) y el salto de línea (\n) constituyen finalizadores de frase ambiguos, por lo que serán final de frase solo cuando vayan seguidos de una palabra que comienza con mayúscula o un iniciador de frase.

5. Los puntos suspensivos (…) no constituyen nunca finalizadores de frase, ni seguidos de mayúscula ni seguidos de minúscula.

Un finalizador de frase (*Sentence Boundary Symbol; SBS*) es cualquier signo de puntuación o salto de línea que marca el final de una oración. En la siguiente tabla se detallan los SBS contemplados por FreeLing3.1 y su frecuencia absoluta y relativa de aparición en el corpus de textos clínicos. La frecuencia relativa se ha calculado en base a un corpus de 1.000.000 palabras:

Sentence Boundary Symbols	***Encoding***	***Frec. Abs.***	***Frec. Rel.***
Punto	*.*	*15962*	*0.038327*
Interrogación de cierre	*?*	*4*	*0.0000096*
Exclamación de cierre	*!*	*0*	*0*
Salto de línea	*\n*	*15445*	*0.0373259*

Puntos suspensivos	*…*	*8*	*0.0000192*

Tabla 1. Finalizadores de frase en FreeLing3.1 baseline

6. Los iniciadores de frase (en todos los casos no ambiguos) son la admiración de apertura (¡) y la interrogación de apertura (¿).

Estas reglas están declaradas en el fichero *splitter.dat*.

3 OTRAS HERRAMIENTAS ANALIZADAS

Las siguientes herramientas se citarán 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 PLN 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 segmentación de frases.

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 segmentación de frases. 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 scripts en Perl centrados en la correcta tokenización de expresiones alfanuméricas propias del ámbito biomédico y asignación del POS para dichas expresiones.

c. Asignación del *Part of Speech* (POS) para el resto del corpus mediante una versión modificada del JunK tagger [6].

d. Proceso de corrección realizado por anotadores humanos.

3.3 GENIA SENTENCE SPLITTER

GeniaSS [12] es una herramienta de segmentación de oraciones implementada para textos biomédicos y entrenada con el corpus GENIA [14]. GeniaSS lee un texto y lo segmenta en oraciones insertando saltos de línea entre las mismas. El modelo de clasificación está basado en métodos de aprendizaje supervisado. Para segmentar correctamente, GeniaSS primero delimita los posibles candidatos a frontera oracional, y después comprueba el contexto donde se encuentran: signos de puntuación colindantes (comas, paréntesis, etc.), primera letra en mayúscula o minúscula de las palabras que preceden y siguen al posible candidato, existencia de mayúsculas en la palabra que le precede, existencia de mayúsculas en la palabra que le sigue, existencia de paréntesis, comillas o comas y apariciones a una distancia de una palabra y a una distancia de dos palabras.

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 segmentación.
- Reglas positivas: Reglas que aplican a casos específicos donde sí se segmentan oraciones. Se acompañan de ejemplos en los que la segmentación se señala con (//).
- Reglas negativas: Reglas que aplican a casos específicos donde no se segmentan oraciones. Si a la regla positiva le aplica una regla negativa se señala en cada caso. Se acompañan de ejemplos.
- Reglas ortográficas: Reglas que aplican a errores de ortotipografía. Se acompañan de ejemplos.

4.1 REGLAS GENERALES (REGLAS-G)

- **G1. Finalizadores de frase**

Un finalizador de frase (SBS) es cualquier signo de puntuación o salto de línea que marca el final de una oración. En el caso de que cualesquiera de los SBS estuviesen repetidos al final de una oración, se etiquetarán por separado considerándose finalizador de frase únicamente el último de ellos.

- Un finalizador de frase ambiguo es aquel que constituye final de frase o no en función de las reglas que se le apliquen.
- Un finalizador de frase no ambiguo es aquel que en cualquier caso siempre constituye un finalizador de frase.

La Tabla 2 recoge los elementos finalizadores de frase potenciales de acuerdo con [16] y añade además otros que han sido detectados tras realizar un estudio detallado del corpus.

SBS	***Encoding***	***Ambiguo/no ambiguo***	***Frec. Abs***	***Frec. Rel***
Punto	*.*	*ambiguo*	*15962*	*0.038*
Interrogación de cierre	*?*	*ambiguo*	*4*	*0.0000096*
Exclamación de cierre	*!*	*ambiguo*	*0*	*0*
Salto de línea	*\n*	*no ambiguo*	*15445*	*0.0373259*
Puntos suspensivos	*…*	*ambiguo*	*8*	*0.0000192*
Coma	*,*	*no ambiguo*	*19180*	*0.0460541*
Dos Puntos	*:*	*ambiguo*	*1404*	*0.0033712*
Paréntesis de cierre	*)*	*ambiguo*	*3710*	*0.008908*
Llave de cierre	*}*	*ambiguo*	*1*	*0.0000024*
Nº total de tokens	*416466*			

Tabla 2. SBS potenciales

Tras el estudio del corpus se ha decidido considerar como finalizadores de frase ambiguos el punto (.), la interrogación de cierre (?), la exclamación de cierre (!), los puntos suspensivos (…), los dos puntos (:), el paréntesis de cierre ()) y la llave de cierre (}). Tanto el salto de línea (\n) como la coma (,) se consideran finalizadores de frase no ambiguos.

CTAKES+UIMA ClinicalPipeline: A continuación se detalla el tratamiento general que CTAKES ClinicalPipeline hace de los SBS potenciales:

- El punto (.) es considerado como un finalizador de frase ambiguo en función de si es detectado como parte de siglas y abreviaturas o no.

- La interrogación de cierre (?) no es considerada un SBS.
- La exclamación de cierre (!) no es considerada un SBS.
- El salto de línea (\n) es un SBS no ambiguo ya que segmenta siempre.
- Los puntos suspensivos (…) son un finalizador de frase ambiguo, ya que segmentan si van seguidos de mayúscula.
- La coma (,) no es considerada un SBS.
- Los dos puntos (:) son un SBS ambiguo. Los dos puntos (:) no segmentan si van entre dígitos ('20:40') pero sí segmentan si van entre palabras o palabras y espacios.
- El paréntesis de cierre ()) es un SBS no ambiguo: segmenta oración siempre.
- La llave de cierre (}) no es considerada un SBS: no segmenta ni seguido de mayúscula ni de minúscula.
- Se considera el salto de línea (\n) como finalizador de frase no ambiguo, ya que segmenta cuando la palabra siguiente empieza tanto con mayúscula como con minúscula.

GENIA (Annotation Guidelines, 2006): No especifica sobre el salto de línea (\n). El punto es siempre finalizador de frase excepto cuando se encuentra entre dígitos, ya que los puntos de abreviaturas, siglas e iniciales se han eliminado del corpus inicial. No especifica en relación al resto de SBS.

GeniaSS: A continuación se detalla el tratamiento general que GeniaSS hace de los SBS potenciales:

- El punto (.) siempre es finalizador de frase a excepción de los casos en los que se encuentra incluido dentro de una sigla o abreviatura reconocida por el sistema o entre dígitos. Es inconsistente con los puntos que aparecen en marcadores o viñetas: en ocasiones los reconoce como parte de las viñetas y en otras ocasiones segmenta oración.
- La exclamación (!) de cierre no se considera un SBS, debido a que no segmenta en ningún caso.
- El salto de línea (\n) se considera SBS tanto cuando la palabra siguiente comienza con mayúscula como cuando la palabra que inicia la oración siguiente comienza con minúscula.
- Los puntos suspensivos (…) son considerados como un SBS no ambiguo, ya que siempre segmentan oración.

- La coma (,) no se considera SBS, ya que no segmenta en ningún caso.
- Los dos puntos (:) no se consideran un SBS. No segmentan en ningún caso.
- Ni el paréntesis de cierre ()) ni la llave de cierre (}) son considerados como SBS. No segmentan en ningún caso.

FreeLing3.1: (Para baseline de reglas de segmentación ver 2.1). Por defecto no considera el salto de línea como finalizador de frase, pero se puede implementar esta funcionalidad (ver implementación I1). Los puntos suspensivos (…) no son considerados en ningún caso como finalizador de frase.

- **G2. Puntos en iniciales, siglas y abreviaturas**

Los problemas de segmentación de frases derivados de los puntos que acompañan a iniciales, siglas y abreviaturas y que no son finalizadores de frase se solucionan mediante reglas de tokenización (iniciales) y normalización (abreviaturas y siglas). Al quedar la sigla o la abreviatura normalizada como un solo token, el punto (.) deja de considerarse finalizador de frase.

CTAKES+UIMA ClinicalPipeline: Incluye algunas siglas entre sus recursos de normalización de modo que estas se anotan como un solo token junto con el punto (.). De esta manera el punto deja de considerarse finalizador de frase. Sin embargo, en otros casos no es consistente, ya que segmenta correctamente solo si reconoce la sigla por aparecer en los corpus de entrenamiento. En español no reconoce las siglas a menos que coincidan con el inglés.

GENIA (Annotation Guidelines, 2006): El punto (.) es siempre finalizador de frase excepto cuando se encuentra entre dígitos, ya que los puntos de abreviaturas, siglas e iniciales se han eliminado del corpus inicial. No especifica sobre puntos que no sean finalizadores de frase.

GeniaSS: Los puntos (.) incluidos en siglas y abreviaturas relativas a dominio biomédico no son considerados finalizador de frase. Sin embargo, cuando se trata de siglas, abreviaturas o unidades de medida no pertenecientes ni al dominio clínico ni al inglés, el punto constituye un finalizador de oración.

FreeLing3.1: Incluye las siglas en un fichero de normalización de modo que se anotan como un solo token junto con el punto (.). De esta manera el punto deja de considerarse finalizador de frase. Este fichero es modificable y se puede enriquecer (ver implementación I2).

- **G3. Revisión de las reglas de anotación**

Si se detectan casos especiales de segmentación de frases no especificados en esta guía se debe reportar el ejemplo para refinar las reglas.

4.2 REGLAS POSITIVAS (REGLAS-P)

- **P1. Salto de línea**

El único caso en el que se reconocerá un final de frase sin punto (.) será si aparece un salto de línea (\n) (Ver regla negativa N1).

Ejemplos:

> *Anticuerpos IgA antitransglutaminasa 0,82 KU/L (0-10). Anticuerpos anti-TG 149 (0-40) y anti-TPO 13,3 (0-35). TSH 3,1 //*
>
> *Exploración ginecológica: útero bicorne.*

- **P2. Punto seguido de espacio + mayúscula**

Si el punto (.) va seguido de una palabra que empieza con mayúscula, se segmenta la frase (ver regla negativa N8).

Ejemplos:

> *Desde que se decidió la apertura de la herida quirúrgica, se hacían lavados con suero fisiológico cada 8 horas y curas con Sulfadiacina argéntica. // El paciente precisó intubación continuada desde el día de la intervención, manteniendo estabilidad hemodinámica gracias al uso de drogas vasoactivas.*
>
> *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. // En la cistoscopia practicada se aprecia un área sobreelevada sin alteración de la mucosa a ese nivel.*

CTAKES+UIMA ClinicalPipeline: Si el punto (.) va seguido de un espacio + mayúscula, siempre se segmenta la frase.

GENIA (Annotation Guidelines, 2006): No especifica.

GeniaSS: Si el punto (.) va seguido de espacio + mayúscula, siempre se segmenta la frase.

FreeLing3.1: Si el punto (.) va seguido de un espacio + mayúscula, siempre se segmenta la frase (ver regla negativa N2).

- **P3. Siglas o abreviaturas con puntos seguidas de mayúscula**

Si el punto (.) que corresponde a la última letra de la sigla o abreviatura va seguido de palabra en mayúscula, se considera final de frase.

Ejemplo:

> […] se inició el tratamiento consistente en la reposición hidroelectrolítica, estabilización hemodinámica mediante el empleo de noradrenalina a dosis de 0,4 mcg/kg/minuto y dobutamina a dosis de 10 mcg/kg/minuto y corrección de la acidosis mediante administración endovenosa de bicarbonato 1 M. // Asimismo, y de acuerdo con el Servicio de Nefrología se decidió la realización de hemodiálisis.

CTAKES+UIMA ClinicalPipeline: Incluye algunas siglas entre sus recursos de normalización de modo que estas se anotan como un solo token junto con el punto (.). De esta manera el punto (.) deja de considerarse finalizador de frase, aunque vaya seguido de mayúscula. Sin embargo, en otros casos no es consistente, ya que segmenta correctamente solo si reconoce la sigla por aparecer en los corpus de entrenamiento. En español no reconoce las siglas a menos que coincidan con el inglés.

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).

GeniaSS: Las siglas o abreviaturas con punto (.) seguidas de mayúscula segmentan oración.

FreeLing3.1: Si el punto (.) que corresponde a la última letra de la sigla va seguido de palabra en mayúscula, se considera final de frase.

- **P4. Unidades de medida con punto seguidas de mayúscula**

Si la unidad de medida se ha escrito con punto (.) al final y va seguida de palabra en mayúscula, siempre se segmenta la frase por considerarse que el punto (.) no es el punto final de la unidad de medida (ver regla ortográfica O1) sino un finalizador de frase. La mayúscula contigua inicia frase.

Ejemplo:

El peso al nacimiento fue 3.480 gramos, la talla 55 cm y perímetro craneal 39,5 cm. // Al nacimiento presenta rasgos dismórficos.

CTAKES+UIMA ClinicalPipeline: Es inconsistente. El punto (.) de la unidad de medida seguida de espacio constituye un finalizador de frase en unos casos y no en otros sin quedar claro el criterio.

GENIA (Annotation Guidelines, 2006): No especifica.

GeniaSS: Las unidades de medida terminadas en punto (.) y seguidas de mayúscula segmentan oración.

FreeLing3.1: El punto (.) de la unidad de medida seguida de espacio y mayúscula constituye un finalizador de frase (ver implementación I3).

- **P5. Puntos suspensivos**

Los puntos suspensivos (…) se considerarán finalizador de frase cuando vayan seguidos de mayúscula.

CTAKES+UIMA ClinicalPipeline: Segmenta oración cuando encuentra puntos suspensivos (…) seguidos de mayúscula.

GENIA (Annotation Guidelines, 2006): No especifica.

GeniaSS: Segmenta oración cuando encuentra puntos suspensivos (…) seguidos de mayúscula.

FreeLing3.1: No segmenta oración cuando encuentra puntos suspensivos (…) seguidos de mayúscula.

4.3 REGLAS NEGATIVAS (REGLAS-N)

- **N1. Finales de frase sin punto**

Si no aparece un punto (.) al final, no se segmenta la frase, independientemente de si la palabra siguiente empieza con mayúscula. (Ver regla negativa N1).

Ejemplo:

El examen histopatológico mostró un tejido fibrosocolagenizado orientado en haces en todos los planos del espacio, sin atipias y revestido por epitelio plano estratificado, compatible con una hiperplasia fibrosa Un mes después de la cirugía se rehabilitó el maxilar superior del paciente mediante una prótesis removible de metal-resina.

CTAKES+UIMA ClinicalPipeline: Si no aparece un punto (.) o un salto de línea (\n), no segmenta la frase. Considera como finalizador de frase tanto el punto (.) como el salto de línea (\n) cuando la palabra siguiente empieza tanto con maýuscula como con minúscula.

GENIA (Annotation Guidelines, 2006): No especifica.

GeniaSS: Si no aparece punto (.) al final de una frase no la segmenta incluso si la palabra por la que comienza la siguiente frase comienza por mayúscula. En el caso de los saltos de línea (\n), GeniaSS segmenta oración siempre independientemente de que la palabra que comienza la oración siguiente comience con mayúscula o con minúscula.

FreeLing3.1: Si no aparece un punto (.) seguido de mayúscula, una admiración de cierre (!) seguida de mayúscula o una interrogación de cierre (?) seguida de mayúscula, no segmenta la frase (ver implementación I1).

- **N2. Punto seguido de espacio + minúscula**

Si el punto (.) va seguido de una palabra que empieza con minúscula, no se segmenta la frase.

Ejemplo:

> *La cantidad de llenado por sesión osciló entre 20 y 60cc. por expansor hasta llegar a un volumen total de 360cc. repartidos entre 7 sesiones y con un intervalo de 3 semanas de media.*

CTAKES+UIMA ClinicalPipeline: Si el punto (.) va seguido de un espacio + minúscula, segmenta la frase inconsistentemente en unos casos sí y en otros no.

GENIA (Annotation Guidelines, 2006): No especifica.

GeniaSS: Es inconsistente en cuanto a la segmentación de punto (.) seguido de espacio + minúscula.

FreeLing3.1: Si el punto (.) va seguido de un espacio + minúscula, no segmenta la frase en ningún caso.

- **N3. Siglas o abreviaturas con puntos seguidas de minúscula**

Si el punto (.) que corresponde a la sigla o abreviatura va seguido de palabra en minúscula, no se segmenta la frase en ningún caso. Esto evita que los puntos (.) que aparecen junto a siglas o abreviaturas supongan un final de frase.

Ejemplo:

> *A las 9:20 a.m. realizamos anestesia raquídea con aguja tipo sprotte (Pajunk®) de 25 G con 2,5ml de bupivacaína 0,5 % (12,5 mg) obteniendo un bloqueo completo sin incidencias.*

CTAKES+UIMA ClinicalPipeline: No es consistente en el tratamiento de las siglas con puntos (.). En ocasiones recoge la sigla como un solo token y en otras la tokeniza con los puntos (.) por separado, por lo que el último punto (.) se convierte en un finalizador de frase y las segmenta.

GENIA (Annotation Guidelines, 2006): No especifica.

GeniaSS: es inconsistente con el tratamiento de de siglas o abreviaturas con puntos (.) seguidas de minúscula. Las abreviaturas que no reconoce ('dr.' 'sr.') segmentan oración. Las abreviaturas que sí reconoce ('Dr.') no segmentan oración.

FreeLing3.1: Si la sigla está recogida en el fichero de normalización (ver implementación I2), considera el punto (.) como parte del token y no lo considera finalizador de frase. FreeLing tampoco segmentará oraciones si la sigla no está recogida en el fichero de normalización y aparece en el corpus seguida de espacio + minúscula.

- **N4. Unidades de medida con punto seguidas de minúscula**

Si la unidad de medida se ha escrito con punto al final (ver regla ortográfica O1) y va seguida de palabra en minúscula, no se segmenta la frase en ningún caso. Esto evita que los puntos (.) que aparecen junto a unidades de medida supongan un final de frase.

Ejemplos:

> *En el estudio anatomopatológico de la pieza se informa como riñón que presenta tumoración de 9x10 cm. multiquística que ocupa la mayor parte del riñón.*
>
> *Se obtuvo muestras de sangre para marcadores testiculares: alfafetoproteína 15000 ng/ml y betaHCG de 200.000 mUI/ml. y se programó orquiectomía para el día siguiente.*

CTAKES+UIMA ClinicalPipeline: El punto (.) de la unidad de medida seguida de espacio constituye un finalizador de frase.

GENIA (Annotation Guidelines, 2006): No especifica.

GeniaSS: Las unidades de medida terminadas en punto ('kg.') y seguidas de minúscula segmentan oración.

FreeLing3.1: El punto (.) de la unidad de medida seguida de espacio y minúscula no constituye en ningún caso un finalizador de frase.

- **N5. Enumeraciones**

El punto (1.) o el paréntesis (a)) de las enumeraciones no se reconocerá como finalizador de frase (Warner et al. 2012: 9).

Ejemplo:

> *Esta propuesta consistía en:*
>
> 1. Resección de las áreas vaginales con heridas de repetición, y cobertura con colgajos locales de rotación.
>
> Se le suministraron dos fármacos: 1. Insulina intravenosa y 2. Insulina humana de acción rápida.

CTAKES+UIMA ClinicalPipeline: No reconoce el punto (.) ni el paréntesis de cierre ()) de las enumeraciones como finalizador de frase, ni cuando va seguido de mayúscula ni cuando va seguido de minúscula.

GENIA (Annotation Guidelines, 2006): No especifica.

GeniaSS: el paréntesis de cierre ()) de las enumeraciones no es identificado como finalizador de frase. El punto (.) de las enumeraciones es tratado de manera inconsistente: detrás de dos puntos (:) segmenta oración mientras que cuando no va precedido por ningún signo de puntuación no segmenta.

FreeLing3.1: Reconoce el punto (.) de las enumeraciones como final de frase pero no el paréntesis de cierre ())(ver implementación I4).

- **N6. Puntos en iniciales**

El punto (.) correspondiente a la inicial de un nombre propio no constituye un finalizador de frase siempre que la inicial esté en mayúscula y el apellido que siga o preceda también comience con mayúscula.

Ejemplo:

> *Se practica adenomectomía retropúbica según técnica de T. Millin, enucleándose un gran adenoma de 170 gramos de peso.*

CTAKES+UIMA ClinicalPipeline: No considera finalizador de frase el punto (.) correspondiente a una inicial de un nombre propio, ni cuando va seguido de mayúscula ni cuando va seguido de minúscula.

GENIA (Annotation Guidelines, 2006): No especifica.

GeniaSS: Considera finalizador de frase el punto (.) que sigue a la inicial del nombre cuando la palabra que sigue va en mayúscula ('L. Hoffman'). No considera finalizador de frase el punto (.) que sigue a una inicial cuando va seguido de minúscula (L. monocytogenes).

FreeLing3.1: No considera finalizador de frase el punto (.) correspondiente a una inicial de un nombre propio ni cuando va seguido de mayúscula ni cuando va seguido de minúscula.

- **N7. Dos puntos**

Los dos puntos (:) no se considerarán final de frase en ningún caso.

Ejemplo:

> Varón de 82 años de edad que acude al Servicio de Urgencias por: disnea, aumento de expectoración y somnolencia.

CTAKES+UIMA ClinicalPipeline: Considera los dos puntos (:) finalizador de frase tanto seguidos de mayúscula como de minúscula.

GENIA (Annotation Guidelines, 2006): No especifica.

GeniaSS: Los dos puntos (:) no se consideran finalizador de frase en ni cuando van seguidos de mayúscula ni de minúscula.

FreeLing3.1: Los dos puntos (:) no se consideran finalizador de frase ni cuando van seguidos de mayúscula ni cuando van seguidos de minúscula. Los dos puntos (:) tampoco constituyen un finalizador de frase cuando van seguidos de salto de línea (\n).

- **N8. Punto seguido de mayúscula**

N8.1 El punto (.) seguido de mayúscula no se considerará finalizador de frase cuando corresponda a la inicial de un nombre propio en mayúscula seguida de uns palabra en mayúscula.

Ejemplo:

Cultivos de semen: E. Coli (1o), S. Sprophiticus (2o), P. Mirabilis (3o).

N8.2 El punto (.) seguido de mayúscula no se considerará finalizador de frase cuando el punto (.) vaya tras un número y la mayúscula corresponda a la primera palabra de una enumeración.

Ejemplos:

1. Inicio de NPT con aporte exclusivo de aminoácidos, glucosa y electrolitos, utilizándose una fuente de aminoácidos distinta a la de la preparación anterior. // Iniciar siempre en horario de mañana con monitorización estrecha de la paciente.

2. Tras 48 h, si no ha habido manifestaciones de reaparición de la alergia, añadir cantidades crecientes de lípidos a la parenteral durante las próximas 72 h.

3. En caso de buena tolerancia, valorar la necesidad de añadir vitaminas y oligoelementos a la mezcla.

- **N9. Puntos suspensivos**

Los tres puntos suspensivos (…) no se considerarán finalizador de frase cuando vayan seguidos de minúscula.

Ejemplo:

Este informe de normalidad absoluta de órganos abdominales… pero cortes efectuados por debajo de sínfisis pubiana a nivel de muslo derecho descubren un gran absceso en la zona de músculos aductores.

CTAKES+UIMA ClinicalPipeline: Segmenta oración cuando encuentra puntos suspensivos (…) seguidos de minúscula.

GENIA (Annotation Guidelines, 2006): No especifica.

GeniaSS: Segmenta oración cuando encuentra puntos suspensivos (…) seguidos de minúscula.

FreeLing3.1: No segmenta oración cuando encuentra puntos suspensivos (…) seguidos de mayúscula.

4.4 REGLAS ORTOGRÁFICAS (REGLAS-O)

- **O1. Unidades de medida**

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

- **O2. Mayúsculas al inicio de frase**

La primera letra de las palabras que inician una oración ha de escribirse con mayúscula.

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. Salto de línea**

Considerar de manera automática el salto de línea como final de frase (tanto ante mayúscula como ante minúscula) se consigue mediante la inclusión de un parámetro (--flush) en la llamada a FreeLing, ausente de la llamada por defecto.

- **I2. Puntos de abreviaturas y siglas**

Las siglas y abreviaturas acompañadas de punto (.) se incluyen en el fichero de normalización singlewords.dat, de modo que se anotan como un solo token junto con el punto (.). De esta manera, el punto (.) deja de considerarse finalizador de frase a menos que la palabra siguiente esté en mayúscula.

- **I3. Unidades de medida con punto seguidas de mayúscula**

Tokenizar por separado el punto (.) es lo más conveniente, ya que son muy escasas en el corpus las apariciones de una unidad de medida con punto (.) seguida de mayúscula. Al conservar la implementación de base de FreeLing3.1, que tokeniza el punto (.) por separado, se minimiza el error en la anotación automática de la segmentación que se produce cuando hay una abreviatura con punto (.) seguida de mayúscula, derivado del incumplimiento de la regla ortográfica O1.

- **I4. Enumeraciones**

Hacer que el punto (.) de las enumeraciones no segmente la frase pondría en riesgo la segmentación correcta de las frases acabadas en un número, por lo que para minimizar el error en la anotación automática de la segmentación es más conveniente no implementar esta regla.

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*

POS	*Part of Speech*
SBS	*Sentence Boundary Symbol*
TA	Traducción Automática