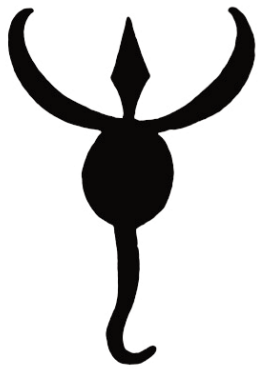
Revista Iberoamericana de Lingüística

RIL



nº 8 2013

R.I.L. 8

Edición, fotomecánica e impresión: Universitas Castellae, edificio 2
Plaza del Viejo Coso, 5
47003 Valladolid
España
Telf. 34 983 377 508 / 629 388 777
E-mail: cuc@universitascastellae.es
www.universitascastellae.es
www.reviblin.com
ISSN: 1887-407X
Fotomecánica e impresión: Universitas Castellae

LA ANOTACIÓN DE LA NEGACIÓN EN UN CORPUS ESCRITO ETIQUETADO SINTÁCTICAMENTE
ANNOTATION OF NEGATION IN A WRITTEN TREEBANK

Antonio Moreno Sandoval y Marta Garrote Salazar
Universidad Autónoma de Madrid

Resumen: En este artículo se describe la anotación de la negación y su ámbito de afectación en un corpus formado por 1501 oraciones extraídas de textos periodísticos y anotadas sintácticamente. En ellas, se han detectado las marcas de negación, los elementos de concordancia negativa y su alcance. Se muestran además los resultados estadísticos preliminares respecto a la frecuencia y funcionamiento de los elementos negativos anotados en el corpus. Este trabajo supone una primera aproximación al tratamiento computacional de la negación. El corpus anotado está disponible de forma gratuita para investigación.
Palabras clave: Anotación de la negación, corpus anotado sintácticamente, marcas de negación, ámbito de la negación.

Abstract: In this paper, it is described the annotation of negation and its scope in a corpus made up of 1501 sentences extracted from newspapers and syntactically annotated. For each sentence, negation cues, negative concordance and their scope have been marked. Also, we show some preliminary statistical results regarding frequency and function of negative elements annotated in the corpus. This work is a first approach to the computational treatment of negation. The annotated corpus is freely available.
Keywords: Negation annotation, treebank, negation cues, negation scope.

1. Introducción

En los últimos años se ha visto un auge de algunas tareas lingüístico-computacionales como la anotación de corpus para el aprendizaje automático conocido como Inferencia Gramatical (Nivre 2006; Seginer 2007). Básicamente, con-

siste en aprender o inferir las reglas gramaticales de una lengua a partir del análisis sintáctico realizado por lingüistas. Tomando la muestra de datos anotados, el programa propone la estructura sintáctica más probable para una oración dada. Esta inducción de reglas gramaticales se realiza mediante métodos estadísticos.

En este trabajo, pretendemos aportar algo de luz sobre uno de los desafíos en la formalización del lenguaje natural: la negación. Mucho se ha escrito sobre ella en el plano teórico de la Lingüística; sin embargo, las teorías no siempre son suficientes en la práctica de formalizar el funcionamiento de la negación para su procesamiento computacional. En los últimos años se ha producido un interés creciente por el tratamiento automático de la negación. El mejor ejemplo de ello es el número especial de la revista *Computational Linguistics* dedicado al tema (Morante y Sporleder 2012), donde se publican cinco artículos que tratan de temas generales aplicados al inglés.

Como explican Morante et al. (2011a:4), "Detecting negated events is important because negated events are not facts". Si un sistema de extracción automática de información se encuentra con la siguiente oración:

(1) La presidenta de Unió Mallorquina no descarta pactar con los socialistas.

debería ser capaz de detectar que el evento principal, *descartar*, está modificado por una marca de negación. Si el sistema extrae que dicha presidenta descarta un pacto con los socialistas, la información que obtenemos es falsa. Para evitar que esto ocurra, el tratamiento de la negación debe ser incorporado al sistema.

Otro ejemplo, del campo de las aplicaciones computacionales en análisis de la opinión, es distinguir entre "Juan lo hace mal", "Juan no lo hace mal" y "Juan no lo hace nada mal". La primera es una valoración negativa mientras que las dos últimas implican una apreciación positiva sobre el sujeto. Las tres contienen elementos negativos.

De entre los trabajos que se están llevando a cabo en este campo, destacaremos algunos por la similitud que tienen con nuestro proyecto: la anotación de la negación y su ámbito en un corpus compuesto por dos historias de Conan Doyle (Morante *et al.* 2011a) y el Bioscope Corpus (Szarvas *et al.*

2008) conjunto de textos médicos en los que se han anotado la negación y su alcance. Ambos corpus están disponibles de forma gratuita para fines de investigación.

Otra muestra de la actualidad del tema son las dos evaluaciones recientes: las tareas de resolución del ámbito y el foco de la negación en *SEM 2012 (Morante y Blanco 2012), donde se anotaron dos conjuntos de datos del PropBank. Otro corpus similar es el *gold standard* elaborado para la tarea *Processing modality and negation* en QA4MRE en el CLEF2012. Este último recurso contiene 8 documentos y 1244 eventos, es decir, en torno a unas 1000 oraciones. Todos ellos son conjuntos de datos en inglés[1].

Este documento, en la línea de los mencionados proyectos, describe una primera etapa del proceso de anotación de información negativa en un corpus de árboles sintácticos, el UAM Spanish Treebank. La principal razón para trabajar con este tipo de corpus y no con otros menos estructurados, mencionados anteriormente es que la anotación sintáctica existente facilita nuestra tarea, especialmente para determinar el ámbito de la negación, y nos permitirá extraer conclusiones más complejas en el futuro. Todas las oraciones y sintagmas negativos han sido anotados, incluyendo tanto las marcas de negación como su ámbito o alcance. El resultado, en formato XML, se puede conseguir de forma gratuita en la dirección http://www.lllf.uam.es/ESP/Treebank.html.

El artículo se estructura de la siguiente manera: en primer lugar, se hace un breve repaso de la situación lingüística teórica de la negación y se expone nuestro modelo de anotación; en segundo lugar, se describe de forma escueta el corpus utilizado para nuestro trabajo; a continuación, se explica el procedimiento de anotación; en cuarto lugar, se muestran algunos de los resultados en términos cuantitativos; por último, se exponen las conclusiones y el trabajo previsto en el futuro. Los ejemplos utilizados pertenecen todos ellos al UAM Spanish Treebank.

2. Marco teórico

La negación se ha tratado en el plano teórico desde perspectivas gramaticales, semántico-lógicas y pragmáticas. Es

precisamente su carácter interdisciplinar lo que la convierte en un elemento lingüístico difícil de formalizar en términos computacionales. Igualmente, ese carácter de la negación la hace computacionalmente inabarcable en un solo paso. Por ello es necesario desglosar su análisis en diferentes fases, desde la puramente formal hasta la más pragmática o dependiente del contexto. Nuestro análisis comienza pues en el nivel gramatical, partiendo de tres problemas fundamentales: las diferentes formas o *manifestaciones lingüísticas básicas* de la negación, la *concordancia negativa* y el *ámbito* de la negación.

Bosque y Gutiérrez-Rexach (2009) destacan la asimetría entre la naturaleza semántica uniforme de la negación y su polivalencia categorial: adverbios (*no*, *tampoco*), pronombres (*nadie*, *nada*), conjunciones (*ni*), preposiciones (*sin*) etc. Todos ellos pueden funcionar como Inductores Negativos (IN), elementos lingüísticos que legitiman a los Términos de Polaridad Negativa (TPN) o palabras y sintagmas que sólo pueden aparecer en entornos negativos (Sánchez 1999, RAE-AALE 2009). En (2), *no* es el IN y *nada* el TPN.

Él *no* hace *nada* por eliminar esa imagen.

Los TPNs son palabras negativas, o palabras-n (*nadie*, *nada*, *ninguno*, *nunca*, *jamás* y *tampoco*) que expresan el significado negativo por sí mismas cuando preceden al verbo. En esta posición, pueden legitimar a otro TPN.

Ayer *nadie* comentó *nada* sobre su estado de salud ni sobre las causas de la enfermedad.

Además, presentan concordancia negativa con el IN, una propiedad morfosintáctica propia de las lenguas románicas (Bosque y Gutiérrez-Rexach, 2009). En (3) no se produce una doble negación, sino la expresión de la negación en dos constituyentes. Esta propiedad diferencia a las palabras-n de aquellas otras palabras con significado negativo intrínseco que combinadas con un IN provocarían una doble negación ("*No* es *im*posible" equivale a la afirmación "Es posible").

Nuestras pautas de anotación contemplan todos los elementos que puedan funcionar como IN y como TPN coordinado con un IN.

Pero la identificación de marcas negativas en un texto no es suficiente: es necesario detectar el *ámbito de la negación*, es decir, el dominio sintáctico sobre el que esta tiene efecto (Sánchez 1999). Sin embargo, el ámbito de la negación es ambiguo

ya que "[…] las interpretaciones de la negación son sumamente flexibles […]" (Bosque y Gutiérrez-Rexach, 2009:638). La decisión más básica es si la marca de negación afecta a la oración entera o solamente a un constituyente.

La teoría gramatical (Sánchez, 1999, Moreno y De Molina 2002, RAE-AALE 2009) habla de *negación externa* o *metalingüística*, que es aquella que abarca toda la oración y *negación interna* o *descriptiva*, que alcanza parcialmente a la proposición o a algunos de sus componentes. Moreno y De Molina, (2002:10) afirman que "tanto la negación como los cuantificadores son operadores que tienen bajo su dominio a los operadores situados a su derecha". Así, se distingue entre negación oracional y sintagmática, la que afecta al sintagma al que precede, o morfológica, la que recae en unidades léxicas mediante prefijos negativos (esta última no se va a tratar en el presente trabajo). Sánchez (1999) también mantiene que el elemento negativo precede a su ámbito y lo domina sintácticamente.

Siguiendo estos supuestos sobre el ámbito de la negación en esta fase del proyecto se van a anotar los siguientes niveles:

1. La negación gramatical (con el inductor "no" o cualquier palabra-n que funcione como inductor en ausencia de este)
    1.1. Oracional: afecta a toda la oración.
        No + (P) + V
    1.2. Sintagmática: afecta sólo al sintagma.
        No + Q/N/ADJ/ADV
2. La negación léxica de las palabras-n
    2.1. Pronombres: *nadie, ninguno, nada.*
    2.2. Adverbios: *nunca, jamás, tampoco.*

Este ha sido el modelo formal empleado para anotar nuestro corpus. Las reglas de 1.1 y 1.2 siguen la teoría sobre la distribución del ámbito de la negación a la derecha del IN. Dependiendo de si a continuación del IN encontramos un verbo o no, la negación será oracional o sintagmática. La negación morfológica mediante prefijos (como *imposible)*, o interna al significado de la palabra (como *rechazar*) y la negación interoracional, representada en su mayoría por conectores oracionales, se tratarán en futuras fases del proyecto. Finalmente, tampoco se tratarán de momento los enunciados negativos sin IN ni TPN: *En lugar alguno, encontrará refugio.*

En resumen, el objeto de esta investigación es centrarse en los aspectos más generales y frecuentes del fenómeno, siempre desde una perspectiva sintáctica oracional. Para ello, se aprovecha el trabajo previo en análisis sintáctico y las herramientas computacionales que permiten localizar y anotar rápidamente las palabras en nuestro corpus. Esto permite realizar un recuento y dar una primera aproximación cuantitativa a estos fenómenos.

3. Descripción del UAM Spanish Treebank

El UAM Spanish Treebank (Moreno *et al.* 2003) es un corpus compuesto por 1501 oraciones anotadas sintácticamente (22695 palabras), todas ellas extraídas de periódicos (*El País Digital* y *Compra Maestra*). Su manual de anotación, disponible junto con el corpus, incluye un inventario de categorías y rasgos, el esquema de anotación e indicaciones específicas sobre una gran variedad de fenómenos del español. Los árboles están codificados en estructura anidada, con los elementos de cada nivel, incluyendo la categoría sintáctica, los rasgos sintácticos y semánticos y los nodos constituyentes, siguiendo el modelo del Penn Treebank. La estructura refleja la sintaxis superficial. La primera fase de este proyecto se desarrolló entre los años 1997 y 2000, y fue el primer corpus anotado sintácticamente del español.

El UAM Spanish Treebank se puede conseguir de forma gratuita, en la dirección http://www.lllf.uam.es/ESP/Treebank.html. Desde 2002, fecha en que se puso a disposición de los investigadores interesados, se han dado más de 40 licencias de uso. De entre las investigaciones más relevantes realizadas con este recurso podemos destacar la llevada a cabo por el grupo de investigación Pattern Recognition and Human Language Technology (PRHLT) en cuya página web (http://prhlt.iti.upv.es/) podemos consultar una herramienta de anotación sintáctica de oraciones. Recientemente, Santamaría (2013) ha demostrado en su tesis doctoral que el UAM Treebank es suficiente para conseguir resultados parecidos a los obtenidos con el PennTreebank con sistemas de inferencia gramatical no supervisada. La Tabla 1 muestra los resultados de la medida armónica F con los sistemas de inducción gramatical CCL (Seginer 2007) y MCM (Santamaría 2013), tomados de la tesis de este último. La evaluación se realizó sobre un sub-

conjunto de los corpus PennTreebank WSJ (inglés) y del UAM Treebank (español) con oraciones de menos de 10 palabras y con el con el corpus completo. El tamaño de los cuatro conjuntos es de 7422 oraciones (WSJ10), 49208 (WSJ-entero), 396 (UAM-10) y 1501 (UAM-entero).

Corpus	10		Completo	
	CCL	MCM	CCL	MCM
WSJ	75	88	56	70
UAM	67	78	59	65

Tabla 1. Evaluación de inferencia gramatical con dos sistemas no supervisados (Santamaría 2013).

Santamaría(2013) muestra que con un conjunto reducido de árboles se pueden obtener resultados bastante similares a los de otro mucho mayor.

Igualmente, llamamos la atención de que el tratamiento de la negación no era ni siquiera un tópico en los corpus sintácticos de primera generación, como se puede comprobar en la monografía de Abeillé (2003).

4. Metodología

La nueva extensión del UAM Spanish Treebank ha sido anotada por dos investigadores expertos, ambos especialistas en Lingüística de Corpus. Para ello, se ha seguido el modelo formal presentado en la sección 2. En los casos de ambigüedad, se ha llegado a un consenso tras una discusión sobre el fenómeno concreto y siempre manteniendo las pautas del modelo establecido[2].

El UAM Spanish Treebank se presenta en formato XML en estructura anidada. Cada una de las oraciones está representada por la etiqueta <Sentence> como elemento raíz dentro del cual se van anidando los diferentes elementos que constituyen la oración. Esto nos permite acotar el ámbito de la negación introduciendo el rasgo Neg="YES" en el constituyente sintáctico sobre el cual recae su significado. Si la negación es total, afectando a la oración completa, el rasgo se incluye en la etiqueta <Sentence>; si por el contrario el ámbito de la negación sólo abarcase, por ejemplo, un sintagma adjetival, representado como <ADJP>, el rasgo Neg="YES" se incluiría en esta etiqueta. El IN o marca de negación se ha marcado con

el rasgo Type=”NEG”, y cuando se produce el fenómeno de concordancia negativa, la palabra en función de TPN se marca con el rasgo Neg=”YES”. Los ejemplos de las Figuras 1, 2 y 3 muestran el resultado de la anotación.

En la Figura 1 vemos un ejemplo de negación oracional (*No comprendía la situación*) en el que se marca el ámbito de la negación (Neg=”YES”) y el IN (Type=”NEG”). La Figura 2 (*No pretende dar ninguna impresión concreta*) se muestra cómo se ha marcado la concordancia negativa: la palabra *ninguna* funciona como TPN concordando con el IN *No* y a su vez actúa como IN del sintagma nominal al que precede. Finalmente, en la Figura 3 (*Carlos Saura y José Luis Garci competirán por la estatuilla a la mejor película de habla no inglesa*) es una muestra de negación cuyo ámbito abarca solamente un sintagma (<ADJP Gender=”FEM” Number=”SG” *Neg=”YES”*>). Los árboles simplificados de las tres oraciones se muestran en la Figura 4. En ellos se puede observar el ámbito de alcance de la negación. Los elementos hermanos del IN (en todos estos casos, el adverbio negativo *no*, excepto en el último NP, cuyo IN es el cuantificador *ninguna*) son núcleos de un constituyente mayor: en el caso del verbo, este es núcleo de la oración; en el caso de otras categorías sintácticas (N, ADJ, etc.), estas son núcleos de un sintagma. Ese constituyente mayor es el alcance de la negación.

```
- <Sentence Neg="YES" Id="129">
    <NP Function="SUBJ" Number="SG" P="3" Elided="Yes"/>
  - <VP Tense="Tensed" Verbal_temp="IMPERFECT" Mode="IND" Number="SG" P="3">
    - <ADVP Type="NEG">
        <ADV Lemma="no" Type="NEG"> No </ADV>
      </ADVP>
      <V Lemma="comprender" Tensed="Yes" Form="IMPERFECT" Mode="IND" Number="SG" P="3"> comprendía </V>
    - <NP Function="OBJ1">
        <ART Lemma="el" Type="DEF" Gender="FEM" Number="SG"> la </ART>
        <N Lemma="situación" Type="Common" Gender="FEM" Number="SG"> situación </N>
      </NP>
    </VP>
    <PUNCT Type="PERIOD"/>
  </Sentence>
```

Figura 1: Ejemplo de negación oracional, *No comprendía la situación*

```
- <Sentence Neg="YES" Id="140">
    <NP Function="SUBJ" Id="1" Gender="SG" P="3" Elided="Yes"/>
  - <VP Tense="Tensed" Verbal_temp="PRES" Mode="IND" Number="SG" P="3">
    - <ADVP Type="NEG">
        <ADV Lemma="no" Type="NEG"> No </ADV>
      </ADVP>
      <V Lemma="pretender" Tensed="Yes" Form="PRES" Mode="IND" Number="SG" P="3"> pretende </V>
    - <CL Type="INFINITIVE" Function="OBJ1">
        <NP Function="SUBJ" Ref="1" Elided="Yes"/>
      - <VP Tense="Untensed" Verbal_temp="INFINITE">
          <V Verbal_temp="dar" Lemma="dar" Tensed="No" Form="INFINITE"> dar </V>
        - <NP Function="OBJ1" Neg="YES">
          - <QP>
              <Q Lemma="ninguno" Gender="FEM" Number="SG" Type="NEG"> ninguna </Q>
            </QP>
            <N Lemma="impresión" Type="Common" Gender="FEM" Number="SG"> impresión </N>
          - <ADJP Gender="FEM" Number="SG">
              <ADJ Lemma="concreto" Gender="FEM" Number="SG"> concreta </ADJ>
            </ADJP>
          </NP>
        </VP>
      </CL>
    </VP>
    <PUNCT Type="PERIOD"/>
  </Sentence>
```

Figura 2: Concordancia negativa en *No pretende dar ninguna impresión concreta*

A diferencia del trabajo de Morante et al. (2011a) y al igual que en el proyecto BioScope (Szarvas et al., 2008), el evento negado (el verbo en el caso de las oraciones o el núcleo en el caso de los sintagmas) no se marca en la actual versión de anotación, aunque se podría incorporar en un futuro. En nuestro corpus, las marcas de negación están incluidas dentro del ámbito de la negación. Todos los argumentos del evento negado se incluyen dentro del ámbito de la negación, incluyendo el sujeto en el caso de las oraciones. La negación afijal y la interoracional no se han tratado en esta fase de anotación, pero ambas están contempladas como proyecto para el futuro.

```
- <Sentence Id="1321">
  - <NP Function="SUBJ" Number="PL" P="3" Coordinated="Yes">
    - <NP Function="SUBJ" Number="SG" P="3">
        <N Lemma="Carlos_Saura" Type="PROPER"> Carlos Saura </N>
      </NP>
      <C Lemma="y" Type="COORDINATING"> y </C>
    - <NP Function="SUBJ" Number="SG" P="3">
        <N Lemma="José_Luis_Garci" Type="PROPER"> José Luis Garci </N>
      </NP>
    </NP>
  - <VP Tense="Tensed" Verbal_temp="FUT" Mode="IND" Number="PL" P="3">
      <V Lemma="competir" Tensed="Yes" Form="FUT" Mode="IND" Number="PL" P="3"> competirán </V>
    - <PP Type="POR" Class="OBL">
        <PREP Lemma="por"> por </PREP>
      - <NP>
          <ART Lemma="el" Type="DEF" Gender="FEM" Number="SG"> la </ART>
          <N Lemma="estatuilla" Type="Common" Gender="FEM" Number="SG"> estatuilla </N>
        - <PP Type="A">
            <PREP Lemma="a"> a </PREP>
          - <NP>
              <ART Lemma="el" Type="DEF" Gender="FEM" Number="SG"> la </ART>
            - <ADJP Type="COMPARATIVE" Gender="FEM" Number="SG">
                <ADJ Lemma="bueno" Type="COMPARATIVE" Gender="FEM" Number="SG"> mejor </ADJ>
              </ADJP>
              <N Lemma="película" Type="Common" Gender="FEM" Number="SG"> película </N>
            - <PP Type="DE">
                <PREP Lemma="de"> de </PREP>
              - <NP>
                  <N Lemma="habla" Type="Common" Gender="FEM" Number="SG"> habla </N>
                - <ADJP Gender="FEM" Number="SG" Neg="YES">
                  - <ADVP Type="NEG">
                      <ADV Lemma="no" Type="NEG"> no </ADV>
                    </ADVP>
                    <ADJ Lemma="inglesa" Gender="FEM" Number="SG"> inglesa </ADJ>
                  </ADJP>
                </NP>
              </PP>
            </NP>
          </PP>
        </NP>
      </PP>
    </VP>
    <PUNCT Type="PERIOD"/>
  </Sentence>
```

Imagen 3

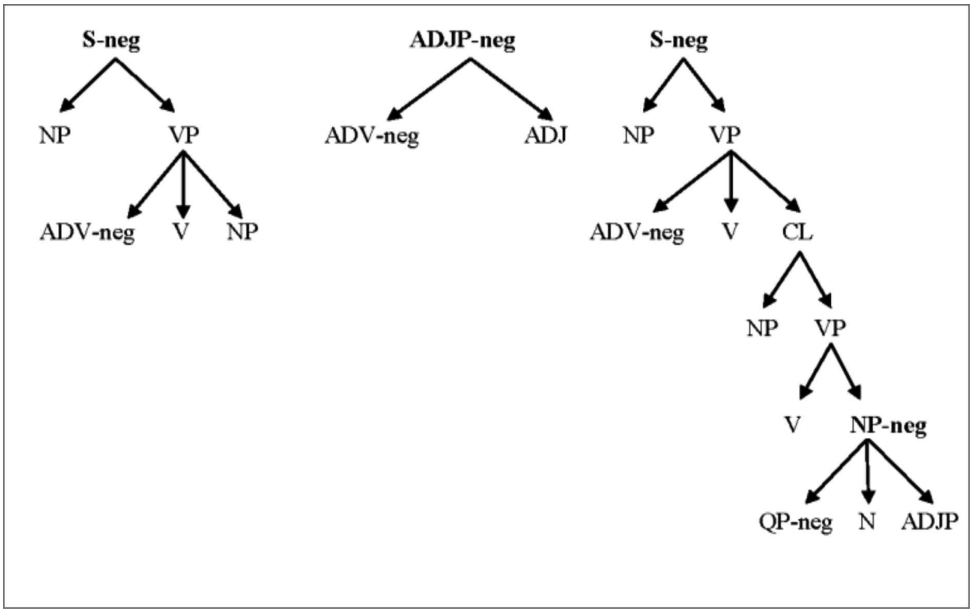


Imagen 4

5. Resultados

Los niveles de la negación anotados en el corpus nos han permitido extraer unos resultados preliminares. El Gráfico 1 muestra el porcentaje que representan las oraciones negativas en el corpus.

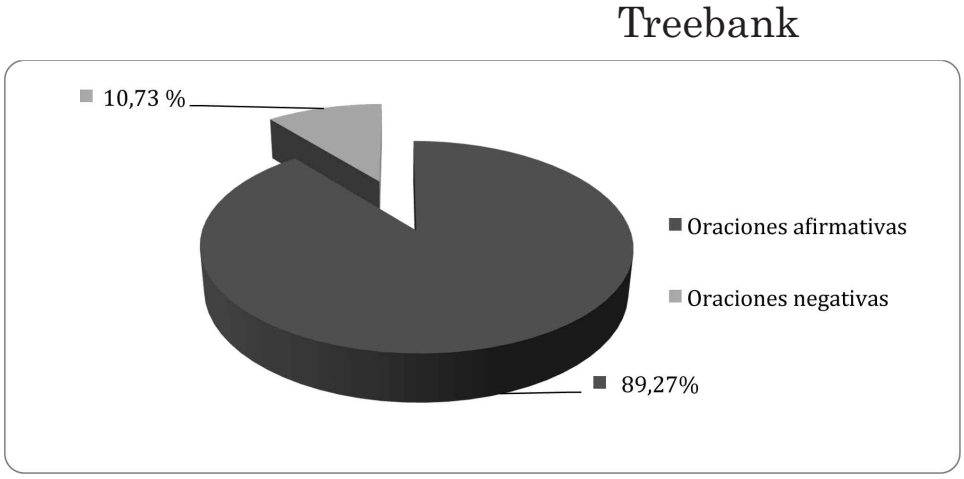


Gráfico 1. Porcentaje de oraciones negativas en el corpus

El hecho de que las oraciones negativas sólo supongan alrededor de un 10%[3] todo el corpus se explica de forma

sencilla porque la negación es un elemento marcado. "En todas las lenguas conocidas, las oraciones negativas se marcan mediante una palabra o morfema especial y las oraciones afirmativas no conocen en general un marcador obligatorio de afirmación." (Moreno Cabrera, 1991: 587). Los elementos lingüísticos marcados son menos frecuentes en las lenguas del mundo por economía lingüística. Por ello, las oraciones afirmativas, y por lo tanto no marcadas, son mucho más productivas en cualquier lengua.

Entre las marcas de negación, el adverbio negativo *no* es la marca más frecuente, un 84,63% del total de las marcas de negación (Tabla 2).

Marcas de negación

Tipo	*Frec. relativa sobre el total del UAM-Treebank*	*Frec. relativa con respecto al total de marcas de negación*
no	9,93	84,63
ni	0,27	2,27
ni_siquiera	0,13	1,14
nunca	0,27	2,27

Tabla 2. Marcas de negación encontradas en el corpus.

Los constituyentes lingüísticos afectados por el ámbito de la negación en nuestro corpus han sido principalmente oraciones, seguidas por sintagmas nominales, adjetivales y preposicionales, todos ellos con una baja frecuencia de afectación (Tabla 3).

Elementos negados	*Frec. total UAM-Treebank*	*Frec. respecto a elementos negados*
Oración	5,13	45,28
Subordinada	5,60	49,39
NP	0,27	2,35
ADJP	0,20	1,76
PP	0,13	1,18

Tabla 3. Elementos lingüísticos afectados por la negación

Las oraciones (ya sean simples o complejas) suponen el 94,70% de total de los elementos negados.

Por último, nos parece significativo el dato sobre la función de las palabras-n (Tabla 4). En nuestro corpus, *nunca* y *tampoco* funcionan en todos los casos de aparición como IN, mientras que *nada* aparece siempre como TPN en concordancia con otro IN. Finalmente, las funciones de *nadie* y *ninguno* están repartidas entre IN y TPN.

Palabras-n	*IN*	*TPN/CN*
ni	100%	
ni_siquiera	100%	
nunca	100%	
nada		100%
nadie	50%	50%
tampoco	100%	
ninguno	25%	75%

Tabla 4. Funcionamiento de las palabras-n

6. Conclusión y trabajo futuro

En este artículo hemos descrito la primera fase del trabajo realizado para anotar la negación y su ámbito. Supone una primera aproximación en español y usando un *treebank*, pues los esfuerzos anteriores se han concentrado en el inglés. Como venimos comentando a lo largo del artículo, la negación se manifiesta en el lenguaje de muy diversas formas y no todas ellas se han tenido en cuenta para este trabajo. que es una contribución desde el campo de la Lingüística de Corpus a la Lingüística descriptiva, por una parte, y a la Lingüística Computacional, por otra. Es incompleto, porque siempre estará limitado por los datos presentes. No podemos incluir toda la tipología de casos que mencionan las gramáticas, simplemente porque para la mayoría de estos no hay ejemplos en las 1501 oraciones del corpus. Nos hemos centrado en lo más básico y

frecuente, la negación sintáctica oracional y sintagmática. La morfoléxica y la discursiva se quedan muy fuera de lo que podemos abordar. Tampoco tratamos la interpretación del foco de la negación y las estructuras negativas que no tienen elementos negativos obvios. La aportación de este artículo es la cuantificación de la negación sintáctica en un corpus equilibrado de oraciones, lo que permite tener una idea aproximada del porcentaje que suponen estas estructuras en el uso lingüístico en textos periodísticos.No obstante, se han diseñado nuevas etapas del proyecto en las que se incluyen la anotación de la negación afijal, la de unidades léxicas intrínsecamente negativas y la negación interoracional. El resultado final será un corpus en el que todos los elementos lingüísticos negativos hayan sido anotados, a partir del cual se puedan extraer patrones de funcionamiento de la negación, con la finalidad de desarrollar sistemas de inferencia gramatical.

Por el momento, podemos concluir que nuestro trabajo demuestra que un corpus sintáctico facilita la identificación de la negación y, sobre todo, de su alcance, gracias a la anotación no sólo de partes del discurso, sino también de elementos superiores como los sintagmas o las oraciones subordinadas. Además, los datos estadísticos nos muestran que solamente tratando el IN "no" y las oraciones como ámbito de alcance se puede abarcar más del 80% de la negación en un texto. A pesar de que las oraciones negativas apenas supongan un 10% del total de oraciones de un texto, su identificación es fundamental para una correcta comprensión del mismo.

Agradecimientos

Esta investigación ha sido financiada por la Comunidad Autónoma de Madrid en el marco del proyecto MA2VICMR (S2009/TIC-1542) y por el MINECO en el marco del proyecto MULTIMEDICA (TIN2010-20644-C03-03).

BIBLIOGRAFÍA

ABEILLÉ, A. (2003). *Treebanks. Building and Using Parsed Corpora*. Netherlands, Kluwer Academic Publishers.

BOSQUE MUÑOZ, I. y GUTIÉRREZ-REXACH, J. (2009). *Fundamentos de sintaxis formal*. Madrid, Akal.

MORANTE, R., SCHRAUWEN, S. y DAELEMANS W. (2011a) "Annotation of negation cues and their scope. Guidelines v1.0". *Technical Report Series CTR-003, CLiPS*. University of Antwerp, Antwerp.

MORANTE, R., SCHRAUWEN, S. y DAELEMANS W. (2011b) "Corpus-based approaches to processing the scope of negation cues: an evaluation of the state of the art" in *Proc. Ninth International Conference on Computational Semantics*, Oxford.

MORANTE, R. y SPORDELER, C. (eds.) (2012) *A Special Issue of the Computational Linguistics Journal on Modality and Negation. Computational Linguistics*, 38:2.

MORENO, A. y DE MOLINA, JA. (2002). *La negación en español*. Granada, Port-Royal.

MORENO CABRERA, JC. (1999) *Curso universitario de lingüística general. Tomo I: Teoría de la gramática y sintaxis general*. Madrid, Síntesis.

MORENO SANDOVAL, A., LÓPEZ RUESGA, S., SÁNCHEZ, F. Y GRISHMAN, R. (2003) "Developing a Spanish Treebank". A. Abeillé (Ed.) *Treebanks. Building and Using Parsed Corpora*. Netherlands, Kluwer Academic Publishers, pp. 149-163.

NIVRE, J. (2006) *Inductive Dependency Parsing*. Dordrecht, Kluwer.

RAE-AALE (2009). *Nueva gramática de la lengua española.* Madrid, Espasa.

SÁNCHEZ LÓPEZ, C. (1999). "La negación". En BOSQUE y DEMONTE (Eds.) *Gramática descriptiva de la lengua española*. Madrid, Espasa Calpe, pp. 2561-2634.

SANTAMARÍA, J. (2013) *Inducción gramatical no supervisada basada en patrones léxicos.* Tesis doctoral inédita. ETSI Informática, UNED.

SEGINER, Y. (2007). *Learning sintactic structure*. PhD Thesis. University of Amstersdam.

SZARVAS, G., VINCZE, V., FARKAS, R. y CSIRIK, J. (2008) "The BioScope corpus: anotation for negation, uncertainty and their scope in biomedical texts". *BioNLP 2008: Current Trends in Biomedical Natural Language Processing*. Columbus, Ohio, pp.38–45.

NOTAS

[1] Morante *et al.* 2011b presentan un breve estado de la cuestión en corpus anotados con la negación, todos ellos recientes y en inglés.

[2] Es habitual en los trabajos de anotación de corpus en los que participan varias personas proporcionar la tasa de acuerdo (*inter-annotator agreement*) para conocer el nivel de certeza o confianza en los datos anotados. En este caso, no procede hablar de ello pues ambos expertos están de acuerdo en el análisis realizado.

[3] En el *dataset* de CLEF12 el porcentaje de eventos negativos es del 5,14 para el inglés. Esto significa que nuestro corpus contiene más información sobre la negación.