Análisis Comparativo de los Sistemas de Anotación de la Negación en Español

Comparative Analysis of the Annotation Guidelines for Negation in Spanish

M. Antònia Martí, Mariona Taulé
CLiC-UBICS, Universitat de Barcelona
Gran Via de Les Corts Catalanes 585,08029 Barcelona
{amarti, mtaule}@ub.edu
http://clic.ub.edu/
http://ubics.ub.edu/

Resumen: En este artículo presentamos un análisis comparativo de las diferentes guías de anotación de la negación en español en el marco de la Tarea-1 del taller NEGES-2018, *Workshop on Negation in Spanish*.
Palabras clave: Negación, guía de anotación, corpus anotados

Abstract: In this paper we present a comparative analysis of the different negation guidelines in Spanish in the framework of Task-1 at the NEGES-2018, *Workshop on Negation in Spanish*.
Keywords: Negation, guidelines, annotated corpora

1 Introducción

En este artículo presentamos un análisis comparativo de las diferentes guías de anotación de la negación en español en el marco de la Tarea 1 del taller NEGES-2018, *Workshop on Negation in Spanish* (Jiménez-Zafra et al., 2018b). El objetivo principal de esta tarea (*Annotation Guidelines*) es analizar los diferentes aspectos relacionados con la negación para llegar a un acuerdo en la definición de unas directrices comunes para la anotación de la negación en español.

El análisis se basa en la comparación de cinco sistemas de anotación, tres de ellos del dominio biomédico, uno sobre opiniones de productos y uno del dominio periodístico. Los tres corpus del dominio biomédico –IULA-SCRC (Marimon, Vivaldi y Bel, 2017), IxaMed-GS (Oronoz et al., 2017) y UHU-HUVR (Cruz et al., 2017)– se caracterizan por contener información semi-estructurada, mientras que los otros dos contienen información no estructurada. El corpus de opiniones –SFU ReviewSP-NEG (Jiménez-Zafra et al., 2018a)– utiliza un lenguaje más informal que el periodístico –UAM Spanish Treebank (Moreno y Garrote, 2013).

El artículo se estructura de la siguiente manera. En la sección 2 se presenta un análisis lingüístico de las características generales de los cinco corpus. En la sección 3 se presentan los datos comparativos de las cinco guías de anotación. En la sección 4 se proponen directrices generales para la anotación de la negación en español como resultado del análisis realizado en 2 y 3. En la sección 5 se presentan las conclusiones.

2 Análisis de los corpus

Una primera distinción a destacar en el análisis comparativo de los 5 corpus presentados en NEGES viene determinada tanto por el dominio temático como por la estructura discursiva de los textos. Estas dos características inciden en el modo de expresar la negación. En este sentido, en el dominio biomédico, las estructuras negativas predominantes constituyen un subconjunto de las estructuras de negación posibles en la lengua: predominan los sintagmas nominales precedidos por un marcador de negación, mayoritariamente 'no' y

Proceedings of NEGES 2018: Workshop on Negation in Spanish, pages 23-28
Seville, Spain, September, 18, 2018

'sin', donde el nombre suele ser una *Named Entity* (NE) del dominio biomédico:

(1) No alergias medicamentosas[1] (UHU-HUVR)
(2) No edemas en extremidades inferiores (IULA-SCRC)

Es importante señalar que las estructuras de (1) y (2) no se atienen a la gramática estándar de la lengua y se pueden considerar características de este dominio. Pueden considerarse oraciones en las que se ha omitido el verbo por ser fácilmente inferible. En general, las oraciones del dominio biomédico son cortas y los predicados verbales utilizados constituyen un conjunto restringido y de alta frecuencia (3), (4) y (5), lo que explica también su omisión.

(3) No ausculto soplos (IULA-SCRC)
(4) Se descarta enolismo (IULA-SCRC)
(5) No visualizamos alteraciones (UHU-HUVR)

Otra característica importante en este dominio es que determinados documentos son textos semiestructurados (por ejemplo, informes radiológicos o anamnesis), donde la negación es un valor de algún tipo de exploración o test médico:

(6) Serología materna: Toxoplasma: Negativo. VHB: Negativo. Rubeola: Negativo. Lues: Negativo. (UHU-HUVR)

En el corpus UHU-HUVR, se anotan como marcadores de negación los símbolos '-' y '/', que tienen este valor solo en determinados dominios. También se utilizan abreviaturas que contienen negación 'namc' (no alergias medicamentos conocidos) como marcadores de negación.

En cuanto a los otros dos corpus analizados en el taller NEGES-2018, aunque el SFU corresponde a un uso más informal de la lengua y el UAM-Treebank es un corpus de la lengua estándar, ambos hacen un uso normativo de la lengua y presentan una estructura discursiva estándar.

[1] Los ejemplos se han obtenido de los artículos en los que se describen los criterios de anotación.

En este caso, fundamentalmente, importa dar cuenta de todas las estructuras de negación posibles y determinar si tienen o no un valor de negación. Se trata de una aproximación que atiende tanto a la cobertura de estructuras posibles como al valor de negación de las expresiones estudiadas. La aproximación que se sigue en ambos casos es de carácter sintáctico y se trata de detectar y anotar todas las posibles estructuras de negación.

Sin embargo, en los corpus de dominio biomédico, el objetivo fundamental es, en último término, determinar si los hechos, eventos o entidades afectados por la negación son o no son factuales: la negación constituye una subtarea en el marco de los sistemas de Extracción de Información (EI) o minería de datos. La negación puede cambiar el estatus factual de la información y, por ello, se sigue una aproximación semántico-pragmática, que adopta soluciones no generalizables ni al dominio general de la lengua ni a otros dominios (7):

(7) Técnicas de Z-N (normal y largo) negativo (UHU-HUVR)

Los autores consideran que no hay negación porque no hay hechos negados ya que la técnica Z-N se ha llevado a cabo (Cruz et al., 2017).

A pesar de estas divergencias debidas al carácter general de dos de los corpus y al carácter específico del dominio biomédico de los otros tres, se puede llevar a cabo una comparación conjunta ya que muchos de los temas abordados son comunes. La comparativa se presenta en la sección 3.

3 Datos comparativos

Los corpus del ámbito biomédico corresponden a informes médicos en soporte electrónico (*Electronic Health Records*, EHRs) del hospital Galdako-Usansolo del País Vasco, como es el caso de IxaMed-GS, y a informes médicos sobre diagnósticos de radiología y de historia personal (anamnesis) del hospital Virgen del Rocío de Sevilla en el caso de UHU-HUVR. En el caso del corpus IULA-SCRC son informes sobre diferentes servicios de un hospital, cuyo nombre no se especifica. El corpus SFU consiste en 400 reseñas de productos, extraídos de la página web Ciao.es y el corpus UAM-Treebank es un corpus formado por oraciones

extraídas de textos periodísticos del UAM-Spanish Treebank (El País Digital y Compra Maestra).

En la Tabla 1 se muestran el número de documentos (Doc.), oraciones y tokens que conforman cada uno de los corpus anotados con negación.

	Doc.	Oraciones	Tokens
IxaMed-GS	75	5.410	41.633
IULA-SCRC	300	1.093	-
UHU-HUVR	604	8.312	145.291
SFU-Review$_{SP}$NEG	400	9.455	221.866
UAM-Treebank	-	1.501	-

Tabla 1: Estadísticas de los corpus

La Tabla 2 resume los distintos aspectos de la negación anotados en cada uno de los corpus: la anotación del marcador de negación (columna 2); el alcance o ámbito (*scope*) de la negación (columna 3); si se incluye o no el sujeto dentro del alcance (columna 4); anotación de estructuras coordinadas (columna 5); anotación de las locuciones de negación (columna 6); anotación de la negación léxica y por afijación (columna 7) y anotación de la especulación (columna 8).

Corpus	Marcador de negación	Scope	Scope (Subj)	Coordinación	Locución Neg.	Nega. Léx./Af.	Especulación
IULA-SCRC	sí	sí	no	sí	-	sí(r)	-
IxaMed-GS	no	sí	-	-	-	no	sí
UHU-HUVR	sí	sí	no	sí	-	sí(r)	-
SFU-Review$_{SP}$-NEG	sí	sí	sí	sí	sí	no	no
UAM-Treebank	sí	sí	sí	-	-	no	no

Tabla 2: Corpus comparativa

Por lo que respecta al marcador de negación, el único corpus que no lo marca explícitamente es el IxaMed-GS. Los otros cuatro sí lo marcan. Los objetivos en función de los cuales se anota un corpus justifican este tipo de decisiones. En el caso de SFU-Review$_{SP}$-NEG y UAM Spanish Treebank, el objetivo de la anotación es obtener datos empíricos sobre las diferentes formas que adopta la negación en español. Por este motivo, se anotan todos aquellos elementos que aportan información relevante sobre el tema, mientras que en IxaMed-GS interesa sólo extraer qué entidades están negadas y, por lo tanto, ignoran el marcador.

Respecto de los marcadores que contienen más de un ítem, SFU-Review$_{SP}$-NEG distingue los continuos (8) y (9) de los discontinuos (10):

(8) La calidad del sonido **no** es mala
(9) **En_mi_vida** he hecho una reserva con tanta antelación
(10) El coche **no** frena **en absoluto**

Mientras que IULA-SCRC y UAM-Treebank para los marcadores complejos, distinguen los ítems que aparecen en posición preverbal (inductores de negación) de los que aparecen en posición postverbal (ítems de polaridad negativa):

(11) **No**$_{[NI]}$ objetivando **ninguna**$_{[NPI]}$ focalidad neurológica (IULA-SCRC)

Todos los corpus han marcado el alcance o ámbito (*scope*) de la negación, pero divergen en si incluyen (SFU-Review$_{SP}$-NEG, UAM-Treebank) o no (UHU-HUVR, IULA-SCRC) el sujeto dentro del ámbito. Consideramos que esta decisión está determinada por las características del dominio biomédico, ya que muchas de las expresiones utilizadas no tienen estructura oracional, por lo que no tiene sentido marcar el sujeto.

Respecto de la negación en estructuras coordinadas, sólo se tratan en SFU-Review$_{SP}$-NEG, UHU-HUVR e IULA-SCRC aunque se abordan de manera distinta. A diferencia de IULA-SCRC y UHU-HUVR, en SFU-Review$_{SP}$-NEG se distingue entre las estructuras negativas coordinadas (12) y las estructuras negativas que contienen elementos coordinados, anotados mediante la etiqueta *discid* (elementos discontinuos) (13). En el primer caso, cada marcador de negación tiene su propio *scope* (12), mientras que en el segundo caso el *scope* incluye toda la coordinación (13).

(12) **No** [soy muy alta] **tampoco** [un pitufo] (SFU-Review$_{SP}$-NEG)
(13) **No** [es ni muy pesado **ni** muy ligero] (SFU-Review$_{SP}$-NEG)

Sin embargo, esta distinción no se establece en los corpus de IULA-SCRC y UHU-HUVR, que adoptan soluciones distintas. En el corpus de IULA-SCRC siempre se incluye la coordinación dentro del *scope* (14), mientras que en UHU-HUVR cualquier tipo de coordinación da lugar a dos estructuras distintas con sus marcadores de negación y correspondientes *scopes* (15) y (16).

(14) **No** [masas ni megalias] (IULA-SCRC)
(15) **No** hemos observado [alteraciones a nivel de los distintos ligamentos (…)], **así como** [de las restantes partes blandas]. (UHU-HUVR)
(16) **No** [hay evidencia de módulos pulmonares] / [adenomegalias mediastínicas] (UHU-HUVR)

El ejemplo (16), según los criterios de SFU, se interpretaría como una sola estructura negativa y, por lo tanto, con un único *scope*. Estas estructuras dependen de la interpretación del anotador, según considere que la estructura coordinada tiene, o no, el mismo predicado verbal, generalmente elidido.

Sólo SFU-Review$_{SP}$-NEG trata explícitamente las locuciones de negación. Éstas incluyen todo tipo de expresión multipalabra que expresa negación, contenga o no un marcador de negación, por ejemplo: ‘en la vida’, ‘en absoluto’.

UHU-HUVR e IULA-SCRC anotan la negación léxica, aunque de manera restringida. IULA-SCRC limita la negación léxica a los predicados: ‘descartar’, ‘ausencia de’, ‘incapaz de’. En el caso de UHU-HUVR, sólo se da unos ejemplos: ‘abandono’, ‘negativo, ‘-‘ y ‘/’, pero no se dan más detalles.

Sólo el corpus IxaMed-GS anota la especulación, tema fundamental en el dominio médico.

4 Propuesta de anotación

Del análisis comparativo de las guías de anotación presentadas en NEGES-2018 se desprende, en una primera aproximación, que el objetivo de la anotación de la negación en los corpus semiestructurados del dominio biomédico y el de los corpus de texto libre presentan marcadas diferencias. Los primeros requieren a menudo una solución *ad hoc* para marcar la negación; en los segundos se aborda la anotación de la negación desde una perspectiva más general, basada en la estructura lingüística.

Ambas aproximaciones son necesarias y adecuadas teniendo en cuenta los objetivos que se quieren alcanzar.

Además, en el dominio biomédico, negación y especulación son dos temas que encontramos tratados conjuntamente en diversos corpus del inglés (Vincze et al. 2008; Vincze, 2010; Konstantinova et al., 2012; Morante y Sporleder, 2012) y también del español (Oronoz et al., 2017). Se trata de dos temas que se interrelacionan ya que ambos inciden en el carácter factual o no factual de lo que se expresa. Este interés que se detecta en un dominio específico, sugiere que se trata de un tema de calado que debería tratarse en un contexto más amplio, en el marco de la lengua en general.

A continuación presentamos nuestras recomendaciones respecto de los diferentes rasgos anotados en los corpus.

Marcadores de negación:

- Siempre que sea posible, consideramos necesario anotar el marcador de negación ya que aporta información al conocimiento general de la lengua: si se marcan, siempre se pueden recuperar y así facilitar, por ejemplo, la creación de un léxico de marcadores de negación.
- Es necesario distinguir los marcadores simples (‘no’, ‘sin’, etc.) de los complejos (‘no…nadie’), donde uno implica la presencia del otro. En este sentido, consideramos pertinente la distinción de SFU-Review$_{SP}$-NEG entre simples y complejos, continuos y discontinuos y, dentro de los complejos, los que actúan como modificadores.

Scope (ámbito o alcance):

Consideramos necesario marcar siempre el *scope* e incluir el sujeto siempre que sea posible. Esta recomendación está justificada por el hecho de que estamos anotando un corpus donde se marca el foco de la negación (Guzzi et al., 2018) y, en algunos casos, éste es el sujeto

(17). Por tanto, no marcar el sujeto dentro del *scope* puede presentar problemas de cara a futuras anotaciones.

(17) Dice que [**no** vendrá *Luisa*].[2]

En (17) el sujeto enfático 'Luisa' está dislocado en posición postverbal y es el foco de la negación, por lo tanto, el sujeto debe incluirse en el ámbito de la negación.

Coordinación:

En el estudio comparativo de las guías, se detectan dos enfoques para el tratamiento de la negación coordinada: a) considerar un único marcador de negación (el primero en el texto) y el resto de la estructura de negación como *scope*, incluyendo los subsiguientes marcadores; y b) distinguir un *scope* para cada marcador de negación coordinado. En SFU-Review$_{SP}$-NEG se contemplan ambas anotaciones, reservando la primera para sintagmas coordinados afectados por un mismo predicado y marcador de negación (13) y la segunda para estructuras coordinadas con marcadores y predicados independientes (12).

Locuciones

En lo que respecta a locuciones, consideramos apropiada la anotación de locuciones que expresan negación, contengan o no (18) marcadores explícitos de negación, ya que en la lengua se da una variedad importante de las mismas y su valor funcional es de negación. En SFU-Review$_{SP}$-NEG se marcan también los marcadores de negación que en determinados contextos no tienen un valor funcional de negación (19).

(18) **En su vida** ha hecho una reserva con tanta antelación.
(19) No pienso irme hasta que **no** vengas.

En el ejemplo (19), el segundo 'no' tiene un valor puramente retórico.

Negación léxica y morfológica

Consideramos también importante identificar la negación léxica y morfológica. Son pocos los corpus que incluyen esta información. En este análisis, sólo se ha aplicado a los corpus UHU-HUVR e IULA-SCRC, aunque de manera restringida. De la observación de los casos anotados, se desprende la necesidad de abordar a fondo el *scope* de la negación léxica y morfológica.

[2] En el ejemplo (17) utilizamos la cursiva para marcar el foco de la negación.

5 Conclusiones

En este trabajo se realiza un análisis comparativo de cinco guías de anotación de la negación en español en el marco de la Tarea 1 del taller NEGES-2018. El análisis parte de la distinción inicial entre corpus pertenecientes al dominio biomédico (textos semiestructurados) y corpus de dominio temático más general como son la noticias y las reseñas de productos, porqué las características del dominio condicionan la estructura discursiva de los textos y la manera de expresar la negación y, en consecuencia, también los criterios de anotación aplicados. Se comparan las distintas propuestas para la anotación de los aspectos básicos tratados en las diferentes guías y se proponen unas directrices (recomendaciones) para la anotación de la negación en español. Las recomendaciones incluyen: 1) la anotación de los marcadores de negación, distinguiendo entre marcadores simples y complejos; 2) la anotación del *scope* o ámbito de la negación, incluyendo el sujeto dentro del ámbito; 3) el tratamiento de la negación coordinada; 4) la anotación de las locuciones negativas (aunque no contengan marcadores explícitos de negación) y 5) la anotación de la negación léxica y morfológica. También sería muy recomendable la anotación del foco de la negación, que no se trata en ninguna de las guías analizadas.

Agradecimientos

Este trabajo ha sido posible gracias al proyecto TIN2015-71147-C2-2 del Ministerio de Economía y Competitividad y a la Generalitat de Catalunya (2017 SGR 3419).

Bibliografía

Cruz, N., R. Morante, M.J. Maña-López, J. Mata-Vázquez y C. L. Parra-Calderón. 2017. Annotating negation in Spanish Clinical Texts. *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, (SemBEaR), páginas 53-58, Valencia, Spain.

Guzzi, E., M.A. Martí, M. Nofre y M. Taulé. 2018. *Guidelines for the annotation of negation in Spanish*, UB, Barcelona.

Jiménez-Zafra, S. M., M. Taulé, M. T. Martín-Valdivia, L. A. Ureña López y M. A. Martí. 2018a. SFU Review$_{SP-NEG}$: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns. *Language, Resources and Evaluation*, 52 (2): 533-569.

Jiménez-Zafra, S. M., N. P. Cruz-Díaz, R. Morante y M. T. Martín-Valdivia. 2018b. Resumen de la Tarea 1 del Taller NEGES 2018: Guías de Anotación. *Proceedings of NEGES 2018: Workshop on Negation in Spanish*, volumen 2174, páginas 15-21.

Konstantinova, N., S. C.M. de Sousa, N.P. Cruz, M.J. Maña, M. Taboada y R. Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. *Proceedings of LREC 2012*. Turquía.

Marimon, M., J. Vivaldi y N. Bel. 2017. Annotation of negation in the IULA Spanish Clinical Record corpus. *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, (SemBEaR), páginas 43-52, Valencia, Spain.

Morante, R. y C. Sporleder. 2012. Modality and Negation: An introduction to the special issue. *Computational Linguistics* 38 (2): 223-260.

Moreno-Sandoval A. y M. Garrote-Salazar. 2013. La anotación de la negación en un corpus escrito etiquetado sintácticamente. *Revista Iberoamericana de Lingüística* 8: 45-60, Valladolid, España.

Oronoz, M., K. Gojenola, A. Pérez, A. Díaz de Ilarraza y A. Casillas. 2015. On the creation of a clinical gold standard corpus in Spanish: Mining adverse drug reactions. *Journal of Biomedical Informatics*, 56: 318-332. Elsevier.

Szarvas, G., V. Vincze, R. Farkas y J. Csirik. 2008. The Bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. *BioNLP 2008: Current Trends in Biomedical Natural Language Processing*, páginas 38-45. Columbus, Ohio. (USA)

Vincze, V. 2010. Speculation and negation annotation in Natural Language Texts: What the case of Bioscope might (not) reveal. *Proceedings of the Workshop on Negation and Speculation in NLP*, ACL páginas 28-31.