Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 39-51 recibido 29-03-2023 revisado 13-05-2023 aceptado 15-05-2023

NoNiRes: Corpus del catalán anotado con negación

NoNiRes: A Catalan corpus annotated with negation

Laura Tañá Velasco[1], Montserrat Nofre Maiz[1], Blanca Calvo Figueras[2], Carme Armentano-Oller[2]
[1] Universitat de Barcelona, CLiC-Centre de Llenguatge i Computació
[2] BSC (Barcelona Supercomputing Center)
{ltanavel, montsenofre}@ub.edu, {blanca.calvo, carme.armentano}@bsc.es

Resumen: En este artículo se presentan los criterios aplicados para la anotación de la negación y del foco de la negación del corpus NoNiRes del catalán. El corpus está constituido por 20.600 oraciones procedentes de *datasets* ya existentes (5.000 oraciones), un foro de Internet (10.000 oraciones) y un periódico digital (5.600 oraciones). Se han tratado aspectos complejos como son el foco y la gradación de la negación. Se ofrecen datos estadísticos exhaustivos sobre las estructuras anotadas.
Palabras clave: Negación, foco de la negación, anotación de corpus, recursos de tecnología lingüística del catalán.

Resumen: In this article we present the criteria applied for the annotation of negation and focus of negation of the corpus NoNiRes of Catalan. The corpus is composed of 20.600 sentences from existing datasets (5.000 sentences), an Internet forum (10.000 sentences), and a digital newspaper (5.600 sentences). Complex aspects such as the focus and the gradation of negation have been dealt with. Comprehensive statistical data on the annotated structures are provided.
Keywords: Negation, focus of negation, corpus annotation, Catalan linguistic resources.

1 Introducción

En las diversas aplicaciones de Procesamiento del Lenguaje Natural (PLN), y muy especialmente en la extracción de información, existe un interés creciente por el estudio y análisis de la negación ya que incide muy directamente en el contenido de la información.

En este artículo se presenta la metodología desarrollada para la constitución de un corpus del catalán anotado con negación, el corpus NoNiRes.

Como apuntan Jiménez-Zafra et al. (2016), la disponibilidad de corpus anotados con negación es esencial tanto para llevar a cabo un estudio teórico de este fenómeno como para el entrenamiento de modelos de inteligencia artificial.

Dado que la forma de expresar la negación es diversa según la lengua de que se trate, se hace imprescindible disponer de recursos específicos para cada una de ellas, puesto que los criterios de detección y anotación de la negación no pueden aplicarse de forma generalizada.

Actualmente, la mayoría de corpus de los que se dispone son del inglés (Jiménez-Zafra et al. 2020; Mahany et al., 2022) aunque también contamos con algunos corpus del español anotados con negación (Jiménez-Zafra et al. 2020). Sin embargo, hasta la fecha no existía ningún corpus del catalán anotado con este tipo de información.

El corpus NoNiRes consta de 20.541 oraciones extraídas de tres tipos de texto: oraciones provenientes de otros conjuntos de datos anotados, textos periodísticos formales y textos generados por usuarios en un foro de Internet (véase sección 3 en este mismo artículo).

ISSN 1135-5948 DOI 10.26342/2023-71-3

Las directrices utilizadas para la anotación se especifican en la *Guia d'Anotació de la Negació,*[1] que se presenta en la sección 4 de este artículo. Dicha guía se ha elaborado a partir de los criterios previamente establecidos en Martí et al. (2016) y Taulé et al. (2021).

En el apartado 2 se presenta un breve estado de la cuestión y en el apartado 3 se expone cómo se ha constituido el corpus.

En el apartado 4, además de revisar las cuestiones generales acerca de la anotación, se analizan los principales problemas encontrados durante el proceso de constitución del corpus. En concreto, se presentan las decisiones adoptadas en lo referente a cuestiones problemáticas como la definición del foco de la negación y las estructuras que, a pesar de contener una partícula negativa, semánticamente no expresan negación. Se han incluido, además, expresiones negativas que no tienen valor de negación y se presenta nuestra propuesta para la resolución de los casos en discusión, con especial atención a las cuestiones referidas a la idiosincrasia del catalán.

En el apartado 5 ofrecemos los resultados obtenidos tras la anotación del corpus y, finalmente, en el apartado 6 proponemos las líneas de trabajo futuras.

2 Estado de la cuestión

2.1 Conceptos preliminares

Previamente a la presentación de los corpus existentes, consideramos necesario esclarecer algunos conceptos básicos que caracterizan el fenómeno de la negación.

En primer lugar, desde un punto de vista lingüístico, se pueden distinguir diferentes tipos de negación: morfológica (derivación: "ilógico"), léxica (palabras de base léxica que por sí mismas expresan negación: "negar", "ignorar") y sintáctica (la que se expresa mediante marcadores sintácticamente libres: "no", "ni").

En segundo lugar, cabe señalar que los sistemas de detección de la negación suelen centrarse en la identificación de los siguientes elementos:

- los marcadores de negación (*negation cues*), es decir, las palabras que expresan la negación.
- el alcance (*scope*) de la negación, es decir, todos los elementos de la oración afectados por la negación.
- el evento negado (*event*), es decir, el elemento directamente afectado por el marcador de negación.
- el foco (*focus*), es decir, el elemento que se niega de forma más prominente o explícita, o la parte del alcance que debe entenderse como negada.

En el corpus NoNiRes se ha anotado la negación sintáctica[2] y, en lo referente a los elementos que la constituyen, el marcador, el alcance y el foco.[3] En el ejemplo 1 aparece una negación sintáctica con dichos elementos marcados.

(1) [**No** hi ha *dret*][4]

"[**No** hay *derecho*]"

Durante el proceso de anotación, para garantizar la consistencia y fiabilidad del resultado final, se han realizado pruebas de acuerdo entre anotadores, tal y como se explica en el apartado 4.1.

Finalmente, en nuestra aproximación al problema utilizamos el concepto de estructura de negación para referirnos a aquellos componentes sintácticos que contienen alguna forma de negación.

2.2 Trabajos previos

Nuestra aproximación a la anotación de la negación se basa en los trabajos realizados anteriormente en el marco del PLN y de la lingüística de corpus. En el caso de los corpus en

1 Tanto las guías de anotación como el corpus están disponibles bajo licencia Creative Commons en la dirección https://zenodo.org/record/7319487.

2 Los mecanismos de la negación sintáctica (mediante partículas de negación independientes) siguen criterios objetivables y claramente definibles (NGLE, 2009). Para la anotación de la negación léxica y morfológica, en la cual intervienen aspectos semánticos, los criterios son más difusos y requieren una fase previa de análisis del problema y de toma de decisiones que no se podía abordar en este trabajo.

3 Para este trabajo, se ha considerado que era más relevante y novedosa la anotación del foco que la del evento. El evento es un rasgo que, inicialmente, ya se abordó en algunos corpus anteriores, aunque no parece una información determinante a la hora de interpretar la negación, mientras que el foco aporta información con mayor impacto en la interpretación del texto (Blanco y Moldovan, 2011).

4 En los ejemplos añadimos una traducción lo más literal posible (T.L.) al español para que se aprecie la estructura en catalán. En algunos casos hemos optado por una traducción un poco más libre (T.C., traducción conceptual). En todos los ejemplos, el marcador se indica en negrita, el foco en cursiva y el alcance entre corchetes.

español, todos estos trabajos siguen los criterios establecidos en la Nueva Gramática de la Lengua Española (NGLE, 2009); en el caso del corpus en catalán, la referencia es la *Gramàtica de la llengua catalana* del Institut d'Estudis Catalans (2016).

En este subapartado presentaremos de forma resumida los corpus anotados con negación en inglés y español. Otros trabajos ya ofrecen una recopilación exhaustiva de este tipo de recursos, e incluyen además corpus en otras lenguas como japonés, chino, alemán o italiano (Jiménez-Zafra et al., 2020; Mahany et al., 2022).

Actualmente, al igual que ocurre en otras tareas de procesamiento del lenguaje, para desarrollar un sistema automático de detección de la negación es necesario disponer de datos de entrenamiento de calidad, es decir, un corpus de un tamaño considerable anotado por expertos que incluya toda la información necesaria para llevar a cabo las tareas propuestas.

Dentro de los corpus existentes anotados con negación, en Jiménez-Zafra et al. (2020) se distingue los que representan la negación de forma lógica, utilizando cuantificadores, predicados y relaciones, y los que enfocan el fenómeno a nivel de cadena, donde el marcador de negación y los distintos elementos (alcance, evento, foco) se definen como partes del texto. En este apartado nos centraremos en estos últimos.

A continuación, mostramos un resumen de los principales corpus de lengua general en inglés y español anotados con negación. Para cada corpus se indica su tamaño, la procedencia de los textos, qué tipo de negación se tiene en cuenta (sintáctica, morfológica, léxica) y qué elementos se han anotado (marcador, alcance, evento, foco).

2.2.1 Corpus anotados con negación (inglés)

- *Product Review corpus*: Formado por 2.111 oraciones procedentes de reseñas de productos extraídas de Google Product Research (679 de ellas contenían expresiones de negación). Se anotaron de forma manual la negación sintáctica: marcadores de negación y alcance (Councill et al., 2010).
- *PropBank Focus* [PB-FOC]: Formado por 3.993 expresiones de negación contenidas en 3.779 oraciones de la sección WSJ del PennTreeBank. Se anotó el foco de la negación (únicamente casos de negación sintáctica), seleccionando la función semántica con mayor probabilidad de ser el foco (Blanco y Moldovan, 2011).
- *ConanDoyle-neg*: Compuesto por 4.423 oraciones procedentes de narraciones escritas por Arthur Conan Doyle, 995 de las cuales contiene negaciones. Se anotaron los marcadores de negación (incluyendo negación sintáctica, morfológica y léxica), el alcance y el evento (Morante y Daelemans, 2012).
- *SFU ReviewEN*: Contiene 17.236 oraciones procedentes de la página web Epinions.com, referidas a diferentes dominios, en 3.017 de las cuales aparece alguna expresión de negación. Se anotaron los marcadores de negación y especulación y, en el caso de la negación, el alcance. En este corpus se anotó la negación sintáctica, pero no la léxica ni la morfológica (Konstantinova et al., 2012).
- *Twitter Negation Corpus*: Contiene 4.000 tuits, 539 de los cuales contienen expresiones de negación (únicamente se tiene en cuenta la negación sintáctica). Se anotaron tanto los marcadores de negación como el alcance (Reitan et al., 2015).
- *Deep Tutor Negation* [DT-Neg]: Formado por 27.785 respuestas de estudiantes a un sistema de diálogo sobre temas de física, 2.603 de las cuales contiene al menos una expresión negativa. Las anotaciones corresponden a los marcadores de negación, el alcance y el foco. Se anotó la negación sintáctica y léxica, pero no la morfológica (Banjade y Rus, 2016).
- *SFU Opinion and Comments* [SOCC]: El corpus original contiene 10.339 artículos de opinión junto con sus 663.173 comentarios, extraídos del diario canadiense The Globe and Mail durante un periodo de cinco años (enero 2012-diciembre 2016), del cual se seleccionaron 1.043 comentarios para anotar el marcador de negación, el foco y el alcance. Se tuvo en cuenta la negación sintáctica, así como algunos verbos y adjetivos que indican la negación. (Kolkhatar et al., 2019)

2.2.2 Corpus anotados con negación (español)

- *UAM Spanish Treebank*: El corpus inicial estaba formado por 1.500 oraciones extraídas de artículos de periódicos, que fueron anotadas sintácticamente. Posteriormente, el corpus se enriqueció con la anotación de la negación: solo 160 oraciones contenían negaciones. Se anotaron los marcadores de negación y el alcance, y se tuvo en cuenta la negación sintáctica, pero no la léxica ni la morfológica (Moreno et al., 2013).
- *SFU ReviewSP-NEG*: Se trata de la versión para el español del corpus SFU Review. Consta de 400 reseñas extraídas de la página web Ciao.es que pertenecen a diferentes dominios; para cada dominio hay 50 opiniones positivas y 50 críticas negativas. El total del corpus lo componen 9.455 oraciones, 3.022 de las cuales contienen al menos un marcador de negación. En este corpus se anotaron los marcadores de negación, el alcance y el evento, y se tuvo en cuenta la negación sintáctica pero no la léxica ni la morfológica (Jiménez-Zafra et al., 2018).
- *NewsCom*: El corpus NewsCom es el primer corpus en español en el que se anota el foco de la negación, en el que se tuvo en cuenta únicamente la negación sintáctica. Consta de 2.955 comentarios publicados en respuesta a 18 artículos de noticias diferentes, obtenidos de periódicos en línea entre agosto de 2017 y mayo 2019, que cubren nueve temas (inmigración, política, tecnología, terrorismo, economía, sociedad, religión, refugiados e inmobiliaria). El 57,80 % de los comentarios (un total de 1.708) contienen al menos una estructura de negación. El número total de estructuras de negación anotadas con foco es de 2.975 (Taule et al., 2021).
- *T-MexNeg*: Corpus en español de México. Consta de 13.704 oraciones procedentes de tuits recogidos entre septiembre de 2017 y abril de 2019, 4.895 de las cuales incluye una negación. Se anotaron los marcadores de negación, el alcance y el evento, solo para casos de negación sintáctica (Bel-Enguix et al., 2021).

Cabe mencionar el corpus NUBes, formado por documentación clínica anonimizada, que consta de 29.682 oraciones, anotadas con negación sintáctica, morfológica y léxica (marcador, alcance y evento) e incertidumbre. En 7.567 de estas oraciones (25,49 %) aparecen estructuras negativas (Lima et al., 2020).

3 Constitución del corpus

3.1 Selección de las oraciones

Para la selección de las oraciones candidatas a ser anotadas hemos usado tres fuentes distintas: a) oraciones procedentes de otros conjuntos de datos anotados, b) oraciones publicadas en un foro de Internet y c) oraciones publicadas en un periódico digital. Llamaremos a estos subconjuntos Datasets, Foros y Periódico, respectivamente. A continuación, describimos cada una de estas fuentes.

3.1.1 Oraciones de otros conjuntos de datos anotados (subconjunto Datasets)

Para la anotación del corpus NoNiRes hemos seleccionado 5.000 oraciones procedentes de dos conjuntos de datos anotados para el catalán ya publicados: el STS-ca, anotado por similitud semántica, y TECa, anotado por implicación textual (Rodriguez-Penagos et al., 2021).[5] Esta selección nos permite tener un mismo conjunto de oraciones con distintos niveles de anotación. La media de tokens de las oraciones procedentes de estos conjuntos de datos es 10,84 tokens, y están escritas en un registro formal.

3.1.2 Oraciones de un foro de Internet (subconjunto Foros)

La segunda fuente de oraciones que hemos usado es Racó Forums,[6] un corpus de textos publicados en el foro de Internet Racó Català, anonimizados.[7]

5 La similitud textual semántica y la implicación textual son tareas que evalúan la relación de significado entre pares de frases. En el primer caso, se mide el grado de similitud semántica entre las dos frases en una escala de 0 a 5. En el segundo caso, se anota la relación lógica entre la primera frase y la segunda, pudiendo ser de implicación, de contradicción o neutra.

6 El corpus Racó Forums está disponible en https://huggingface.co/datasets/projecte-aina/raco_forums

7 https://www.racocatala.cat/forums

Estos textos recogen el uso del idioma propio de los contenidos generados por usuario, con numerosas expresiones coloquiales o, incluso, vulgares, errores ortográficos y construcciones agramaticales.

En concreto, hemos seleccionado 10.000 oraciones de este corpus. Para obtener oraciones de características parecidas a las del grupo anterior, hemos limitado la selección a las oraciones cuya longitud oscila entre 4 y 22 tokens. De media, estas oraciones tienen 11,08 tokens.

3.1.3 Oraciones publicadas en un periódico digital (subconjunto Periódico)

El resto de las oraciones (5.600) del corpus NoNiRes se seleccionaron a partir de noticias del periódico digital VilaWeb.[8]

Para asegurar que en el corpus hubiera un mínimo de oraciones negativas, se seleccionaron 2.800 de entre las que contenían alguno de los marcadores de negación identificados en el grupo de oraciones procedentes de otros conjuntos de datos anotados.[9]

El resto de las oraciones fueron seleccionadas al azar, sin determinar una longitud máxima, aunque, para evitar la selección de titulares que podrían ser oraciones incompletas, se eligieron oraciones con un mínimo de 4 tokens. Como resultado, las oraciones de esta fuente son más largas, con una longitud máxima de 142 tokens, y una media de 27,65 tokens. Como el primer conjunto, el registro lingüístico de estas oraciones es formal.

4 Anotación del corpus

Como ya se ha explicado en el apartado 1, en el corpus NoNiRes se ha anotado la negación sintáctica y, en lo referente a los elementos que la constituyen, el marcador, el alcance y el foco.

Para la anotación del corpus NoNiRes se ha usado la plataforma Prodigy, que permite hacer una anotación manual de spans. Mediante las etiquetas <CUE>, <SCOPE> y <FOCUS> se ha anotado el marcador de negación, el alcance y el foco.

A continuación, explicamos el proceso de anotación y detallamos las cuestiones centrales de las guías de anotación.

4.1 Acuerdo entre anotadores

Con el objetivo de comprobar el funcionamiento de la plataforma escogida para la tarea de anotación y verificar la aplicación de los criterios establecidos, en la fase inicial se anotó en paralelo un pequeño subconjunto de 50 frases, a cargo de dos anotadores expertos (el porcentaje de acuerdo para esta pequeña muestra fue del 87 %, k= 0,67).

Tras este proceso inicial, se actualizó la guía de anotación.

El grueso de la anotación corrió a cargo de un anotador experto principal. Los casos conflictivos o dudosos se discutían de forma regular por el anotador principal, el segundo anotador experto y una lingüista especialista en PLN y anotación de corpus.

Cabe destacar que, tras la primera fase, la guía ya no tuvo que ser modificada durante el resto del proceso de anotación, lo que corrobora la adecuación de los criterios establecidos para la tarea.

4.2 Criterios de anotación

Como se ha expuesto en el apartado 2, nos hemos centrado en la anotación de la negación sintáctica y de los diferentes elementos que conforman la negación: hemos anotado el marcador de negación (*cue*), el alcance (*scope*) y el foco (*focus*).

En nuestra aproximación, entendemos por marcador de negación aquella palabra o grupo de palabras que identifica que estamos ante un hecho o afirmación negados. En general, pertenecen a categorías cerradas como son los adverbios, los determinantes y los pronombres, pero es necesario tener en cuenta que también hay marcadores adverbiales de base léxica. Por ejemplo, la expresión “en ma vida” (“en toda mi vida”) funciona como un adverbio y es equivalente al marcador adverbial “mai” (“nunca”), pero no contiene ningún indicio de expresión de la negación.

Son ejemplos de marcadores de negación en catalán los adverbios “no” y “mai” (“nunca”), el pronombre “ningú” (“nadie”), la conjunción “ni” y el determinante “cap” (“ningún”), entre otros.

8 Periódico digital VilaWeb: https://www.vilaweb.cat/

9 Los marcadores identificados fueron: "cap", "cap més", "encara menys", "enlloc", "excepte", "gaire", "gens", "mai", "menys", "molt", "ni", "ningú", "no", "no només", "només", "pas", "re", "res", "res més", "sense", "tampoc", "tan", "tret".

Hay que tener en cuenta que se dan casos de marcadores discontinuos que forman parte de una sola estructura de negación (ejemplo 2).

(2) [**No** hi ha **cap** *problema*]

"[**No** hay **ningún** *problema*]"

El alcance se suele definir como la máxima unidad sintáctica que se encuentra afectada por el marcador de negación (Morante y Sporleder, 2012). Se trata de una definición vaga que se suele interpretar de manera libre, por lo cual en los corpus puede estar representado de diferentes formas. Existen dos tendencias en la anotación del alcance: una que incluye el marcador de negación y el sujeto del predicado verbal afectado por la negación, y otra que excluye dichos elementos.

En nuestro sistema de anotación hemos incluido dentro del alcance tanto el marcador como el sujeto del predicado verbal negado (subrayados en el ejemplo 3).

(3) [Ells **tampoc** ho fan *amb nosaltres*]

"[Ellos **tampoco** lo hacen *con nosotros*]"

El foco es la parte del alcance más explícita o prominentemente negada (Huddleston y Pullum, 2002). Se trata de un elemento especialmente difícil de identificar, ya que no siempre se dispone de indicadores lingüísticos formalmente explicitados para su detección. Puede estar determinado por la semántica de las palabras o las intenciones comunicativas, y depender de información paralingüística.

En el corpus NoNiRes, para la anotación del foco, hemos partido de los criterios de Taulé et al. (2021) aplicados al corpus NewsCom, el primer y único corpus en español que anota el foco.

4.3 El foco de la negación en catalán

Puesto que el foco de la negación es un componente especialmente difícil de identificar, lo tratamos de manera específica en este apartado.

Para la identificación del foco de la negación partimos del criterio del elemento más oblicuo, según el cual el candidato más probable para ser interpretado como foco es el constituyente más oblicuo, de acuerdo con Taulé et al. (2021). Esta decisión se fundamenta en el hecho de que el elemento más oblicuo explicita información no requerida sintácticamente por el predicado, es decir, no constituye un argumento verbal, de modo que, si no tuviese un valor informativo relevante, no se explicitaría.

Generalmente el elemento oblicuo es un complemento circunstancial o un adjunto oracional (ejemplo 4).

(4) Però [*tàcticament* **no** ho eren tant].

"Pero [*tácticamente* **no** lo eran tanto]."

4.3.1 Casos con más de un candidato a foco

Es frecuente que en una misma oración nos encontremos con más de un adjunto y por lo tanto con más de un candidato a foco de la negación. En estos casos uno de los dos adjuntos suele ser una expresión temporal (subrayada en el ejemplo 5). Cuando esto sucede se genera una ambigüedad, ya que hay más de un elemento susceptible de ser interpretado como foco de la negación.

En estos casos, hemos decidido considerar que la expresión temporal define el marco temporal en el que se realiza el foco de la negación y, por tanto, hemos anotado como foco el otro adjunto oracional, de acuerdo con Tañà (2021).

(5) [De vegades, *en el bosc* **no** havia escombres].

"[A veces, *en el bosque* no había *escobas*]."

4.3.2 Foco elíptico

Existen oraciones (ejemplo 6) que contienen un marcador de negación cuyo foco se halla en una oración anterior que, dadas las características del corpus, no es recuperable. En estos casos, consideramos que el foco es elíptico y que el alcance está formado únicamente por el marcador de negación.

(6) [**No**], tu amb ella i jo amb ell

"[**No**], tú con ella y yo con él

En el ejemplo, se sobreentiende que "no" responde a una intervención anterior que no podemos recuperar y es la que contiene el foco de negación.

4.4 Casos especiales

Incluimos en este apartado las oraciones con gradación de la negación, el uso del adverbio "pas" y aquellas estructuras que, a pesar de contener una partícula negativa, semánticamente no expresan negación (subrayadas en los ejemplos 13 a 19).

En el corpus NoNiRes, este tipo de construcciones son especialmente frecuentes en las oraciones procedentes de foros de Internet.

4.4.1 Gradación de la negación

Algunas oraciones (ejemplos 7, 8 y 9) contienen un marcador de negación (por ejemplo, el adverbio "no") y un segundo elemento que expresa una gradación de la negación, es decir, que la potencia o atenúa.

(7) Sort que [*ja* **no** hi havia **gaire** gent]...

"Suerte que [*ya* **no** había **mucha** gente]..."

(8) [Encendre un foc **no** hauria de ser **gaire** *difícil*].

"[Encender un fuego **no** debería ser **muy** *difícil*]."

(9) [Aquest tercer treball **no** tingué una crítica **tan** *bona* com els anteriors]

"[Este tercer trabajo **no** tuvo una crítica **tan** *buena* como los anteriores]"

"Gaire" y "tan" son cuantificadores que indican un grado no absoluto de negación. Cuando aparecen, hemos optado por anotarlos como marcadores discontinuos de negación y considerar como foco de la negación el elemento que queda afectado.

4.4.2 Uso de "pas"

Un aspecto característico del catalán es el uso del adverbio "pas" como componente de las expresiones negativas (ejemplos 10 y 11). Su función es la de reforzar la negación. En la actualidad, su uso, especialmente en la lengua oral, ha decrecido.

(10) [**No** hem de pensar **pas** *que Déu ens castiga*].

"[**No** debemos pensar **Ø** *que Dios nos castiga*]."

(11) Però [*en qüestions d'amor*, **no** escullo **pas** bàndols].

"Pero [*en cuestiones de amor*, **no** elijo **Ø** bandos]."

En nuestro caso, hemos anotado estos casos como marcadores de negación discontinuos, de forma similar a la que se anotan en corpus en francés, como en el ejemplo 12, donde "n'" y "pas" se consideran un marcador discontinuo. (Seminck, s.f.).

(12) Ils **n'**ont **pas** conduit jusqu'au pont

4.4.3 Expresiones lexicalizadas

Se trata de expresiones que contienen una partícula negativa y que se han lexicalizado, con lo cual han perdido el valor negativo y han adquirido otro tipo de significado (ejemplos 13 y 14, donde se subrayan dichas expresiones).

(13) Acaben dient <u>si fa o no fa</u> el mateix

"Acaban diciendo <u>más o menos</u> lo mismo"

(14) El referèndum el vol convocar <u>tan sí com no</u>

"El referéndum lo quiere convocar <u>sí o sí</u>"

4.4.4 Partículas expletivas de negación

Se trata de partículas que originariamente expresan negación pero que en el contexto en el que aparecen no tienen ningún valor semántico. Son formas expletivas redundantes (en el ejemplo 15 se subraya uno de dichos casos).

(15) RCat treurà més vots que no pas SCI

"RCat sacará más votos que Ø SCI"

4.4.5 Expresiones con valor contrastivo o de contraposición

Se trata de expresiones que introducen una corrección, añaden información nueva o contraponen respecto de un límite (ejemplo 16, subrayado).

(16) És una pocasoltada que no té cap més objectiu que crear confusió.

"Es una tontería que no tiene Ø otro objetivo que crear confusión"

4.4.6 El uso retórico discursivo del lenguaje

En estos casos la partícula negativa (subrayada en el ejemplo 17) se usa con un valor pragmático/comunicativo.

(17) T'agrada la pel·lícula, no?

"Te gusta la película, ¿no?"

4.4.7 Doble negación

El fenómeno de la doble negación tiene lugar cuando en una misma oración aparecen dos o más palabras que expresan negación (Espinal y Tubau, 2010).

Generalmente encontramos un adverbio de negación y un verbo con contenido semántico negativo que cancela el valor del adverbio (por ejemplo, 'negar' y 'descartar'). En los ejemplos 18 y 19 se subrayan algunos casos.

(18) Ningú no podrà negar que és original.

"Nadie Ø podrá negar que es original."

(19) Els Mossos d'Esquadra no descarten més detencions.

"Los Mossos d'Esquadra no descartan más detenciones."

5 Resultados

El corpus NoNiRes consta de 20.541 oraciones anotadas,[10] que incluyen 5.182 estructuras de negación contenidas en 4.488 oraciones, lo que representa un 21,85 % del total de las oraciones del corpus. De estas 4.488 oraciones, 3.881 contienen una sola estructura negativa, 532 contienen dos estructuras de negación, y las 75 restantes contienen más de dos estructuras negativas.

El 21,73 % de los marcadores de negación detectados son discontinuos, como en el ejemplo 20.

(20) **No** tinc casa **ni** cotxe

"**No** tengo casa **ni** coche"

En cuanto al alcance, este incluye, de media, el 50,31 % de tokens de las frases anotadas (8,56 tokens de media). Por lo que refiere al foco, se ha anotado en un 98 % de aparición de marcadores de negación, y la distancia media entre el marcador y el foco es de 1,65 tokens.

Los marcadores de negación más usados en el corpus son "no" (3.302 ocurrencias), "sense" ("sin", 328 ocurrencias), "no [...] cap" ("no [...] ninguno/a", 239 ocurrencias), "no [...] res" ("no [...] nada", 154 ocurrencias) y "ni" (112 ocurrencias). En el Anexo 1 se muestra la lista de los marcadores más usados.

Dado que nuestro corpus contiene tres subconjuntos de datos tomados de diferentes fuentes, conviene observar los resultados de estas tres fuentes por separado. Las Tablas 1 y 2 contienen las estadísticas para cada una de dichas fuentes.

En primer lugar, podemos comparar las oraciones procedentes de Datasets y de Foros, ya que son oraciones de una longitud similar, que se han obtenido sin realizar ningún filtro por marcadores, pero de registros lingüísticos distintos (formal frente a generado por usuario). En estos conjuntos observamos que las oraciones que contienen alguna estructura negativa suponen, proporcionalmente, más del doble en las oraciones de Foros (21,6 %) que en las de Datasets (8,15 %). Esto se acentúa todavía más en las oraciones que contienen dos estructuras negativas, cuatro veces más frecuentes en las oraciones procedentes de foros (2,67 %) que en

10 En el proceso de anotación se descartaron 59 frases que resultaban incomprensibles.

las procedentes de conjuntos de datos (0,66 %). Finalmente, mientras que en las oraciones de Datasets solo hay una oración con más de dos estructuras negativas, en las oraciones de Foros encontramos hasta 23 ejemplos.

Aunque, en principio, las oraciones del subconjunto Periódico deberían tener características parecidas a las del subconjunto Datasets, ya que ambos contienen lenguaje formal, no se pueden comparar en términos de cantidad de estructuras negativas, ya que se trata de oraciones considerablemente más extensas y a las que se ha aplicado un filtro previo por marcadores. No obstante, sí podemos comparar algunos rasgos de sus estructuras negativas. Mientras que en las oraciones de lenguaje formal (Datasets y Periódico) la distancia media entre foco y marcador es de 1,80 y 1,96 tokens, respectivamente, en las oraciones de Foros, generadas informalmente por usuarios, la distancia media entre foco y marcador se reduce a 1,32 tokens. También es reseñable la diferencia en cuanto al número de estructuras negativas con foco elíptico: en Foros representan un 2,82 % del total de estructuras negativas, mientras que en Datasets y Periódico suponen el 1,05 % y el 1,40 % respectivamente.

En cuanto al uso de marcadores, destacan algunas diferencias entre los subconjuntos de lenguaje formal y el de Foros. Por un lado, observamos que el uso de "sense" ("sin") es más frecuente en el lenguaje formal (3,34 % de las estructuras negativas en Foros frente a un 8,4 % y un 8,97 % en Datasets y Periódico, respectivamente). Por el contrario, el marcador "ni" es más frecuente en el lenguaje generado por usuarios (3,18 % de las estructuras negativas en Foros frente 0,63 % y 1,40 % en Datasets y Periódico, respectivamente). Las listas de los marcadores más frecuentes por fuente pueden consultarse en los anexos 2, 3 y 4.

A continuación, se resumen todos estos datos en las Tablas 1 y 2.[11]

	Total	**Datasets**	**Foros**	**Periód**
Orac.	20.541	4.993	9.990	5.558
Tokens/ oración (media)	15,50	10,84	11,08	27,65
Orac. con alguna estr. neg. (%)	4.488 (21,85)	441 (8,83)	2.104 (21,06)	1.943 (34,95)
1 estr. neg. (%)	3.881 (18,8)	407 (8,15)	1.814 (18,1)	1.660 (29,8)
2 estr. neg. (%)	532 (2,59)	33 (0,66)	267 (2,67)	232 (4,17)
>2 estr. neg. (%)	75 (0,37)	1 (0,02)	23 (0,23)	51 (0,92)

Tabla 1: Datos referentes a las oraciones del corpus por fuente de procedencia.

	Todas	**Datasets**	**Foros**	**Periód**
Estr. Neg.	5.182	476	2.420	2.286
Discont. (%)	1.126 (21,73)	110 (23,11)	526 (21,74)	490 (21,43)
Tokens alcance (%)	8,56 (50,31)	6,98 (64,70)	6,84 (56,71)	10,71 (40,53)
Foco elíptico (%)	105 (2,02)	5 (1,05)	68 (2,81)	32 (1,40)
Distancia media (tokens) foco/ marcador	1,65	1,80	1,32	1,96

Tabla 2: Datos referentes a las estructuras negativas del corpus por fuente de procedencia.

11 En el Anexo 5 se ofrece una comparación con los datos obtenidos a partir del corpus NewsCom.

6 Conclusiones y futuras líneas de trabajo

En este artículo hemos presentado los criterios para la anotación de la negación en catalán partiendo de Martí et al. (2016) y Taulé et al., (2021). Nos hemos centrado principalmente en el foco de la negación, puesto que se trata de un elemento cuya identificación resulta difícil, ya que puede estar determinado por la semántica de las palabras y las intenciones comunicativas, y depender de información extralingüística.

Hemos tratado también la gradación de la negación, el uso típico del catalán del adverbio "pas" y diversos casos de oraciones que, a pesar de contener una partícula negativa, no expresan negación.

El corpus NoNiRes es, por lo tanto, el primer corpus del catalán anotado con negación que incluye el marcador, el foco y el alcance. El corpus está formado por 20.541 oraciones, un 21,65 % de las cuales contienen estructuras negativas. La mitad de las oraciones corresponden a textos generados por usuarios en foros (subconjunto Foros), un cuarto procede de conjunto de datos ya existentes (Datasets), y otro cuarto de las oraciones se han obtenido de un periódico digital (Periódico). Comparando los dos primeros conjuntos, hemos observado que las estructuras de negación son más abundantes en los textos generados por usuarios.

NoNiRes constituye un recurso lingüístico interesante para estudiar el uso de la negación en los distintos dominios del lenguaje. Como futuras líneas de trabajo podemos apuntar el uso de este corpus para entrenar y evaluar modelos de detección automática de la negación. Adicionalmente, también proponemos estudiar la relación entre negación y factualidad/contrafactualidad, como en el caso de los enunciados no declarativos expresados mediante una negación.

Agradecimientos

Agradecemos a Racó Català la cesión de sus datos, que han sido usados para la recolección de frases generadas por usuarios.

Agradecemos la implicación y el apoyo de la profesora M. Antònia Martí (Universitat de Barcelona, CLiC-Centre de Llenguatge i Computació) durante todo el proceso de anotación y de elaboración del artículo.

Agradecemos también el asesoramiento de Carlos Gerardo Rodriguez Penagos y Ona de Gibert (BSC-Barcelona Supercomputing Center).

Este trabajo ha sido financiado por CLiC, Centre de Llenguatge i Computació, grupo de investigación consolidado por la Generalitat de Catalunya (2021 SGR 00313), y por el Departament de la Vicepresidència i de Polítiques Digitals i Territori de la Generalitat de Catalunya, dentro del marco del Projecte AINA.

Bibliografía

Armengol-Estapé, J., C. Pío Carrino, C. Rodriguez-Penagos, O. de Gibert Bonet, C. Armentano-Oller, A. Gonzalez-Agirre, M. Melero, y V. Villegas 2021. Are Multilingual Models the Best Choice for Moderately Under-resourced Languages? A Comprehensive Assessment for Catalan. *Findings of the Association for Computational Linguistics: ACL-IJCNLP* 2021, 4933–4946.

Banjade, R., y V. Rus. 2016. DT-Neg: Tutorial dialogues annotated for negation scope and focus in context. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 3768–3771.

Bel-Enguix, G., H. Gómez-Adorno, A. Pimentel, S.-L. Ojeda-Trueba, y B. Aguilar-Vizuet. Negation Detection on Mexican Spanish Tweets: The T-MexNeg Corpus. *Applied Sciences*, 11: 3880.

Blanco, E., y D. Moldovan. 2011. Semantic representation of negation using focus detection. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 581–589.

Councill, I. G., R. McDonald, y L. Velikovich. 2010. What's great and what's not: Learning to classify the scope of negation for improved sentiment analysis. *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, 51–59.

Gramàtica de la llengua catalana. 2016. Barcelona: Institut d'Estudis Catalans.

Jiménez-Zafra, S. M., M. T. Martín-Valdivia, L. A. Ureña López, M. A. Martí, y M. Taulé. 2016. Problematic Cases in the Annotation of Negation in Spanish. *Proceedings of the Workshop Extra-Propositional Aspects of*

Meaning in Computational Linguistics (ExProM-2016): 45-48.

Jiménez-Zafra, S. M., R. Morante, M. T. Martín-Valdivia, y L. A. Ureña-López. 2018. A review of Spanish corpora annotated with negation. *Proceedings of the 27th International Conference on Computational Linguistics*, 915–924.

Jiménez-Zafra, S. M., M. Taulé, M. T. Martín-Valdivia, L. A. Ureña-López, y M. A. Martí. 2018. SFU ReviewSP-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., R. Morante, M. T. Martín-Valdivia, y L. A. Ureña-López. 2020. Corpora Annotated with Negation: An Overview. *Computational Linguistics*, 46(1):1–52.

Kolhatkar, V., H. Wu, L. Cavasso, E. Francis, K. Shukla, y M. Taboada. 2019. The SFU opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 1–36.

Konstantinova, N., S. C. M. De Sousa, N. P. Díaz Cruz, M. J. Maña Lopez, M. Taboada, y R. Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 3190–3195.

Lima López, S., Pérez, N., Cuadros, M. y Rigau. G. 2020. NUBes: A Corpus of Negation and Uncertainty in Spanish Clinical Texts. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 5772–5781.

Mahany, A., H. Khaled, N. Sabri, N. Aljohani, y S. Ghoniemy. 2022. Negation and Speculation in NLP: A Survey, Corpora, Methods, and Applications. *Applied Sciences*. 12: 5209.

Martí, M. A., M. T. Martín-Valdivia, M. Taulé, S. M. Jiménez-Zafra, M. Nofre, y L. Marsó. 2016. La negación en español: análisis y tipología de patrones de negación. *Procesamiento del Lenguaje Natural*, 57: 41-48.

Morante, R., y W. Daelemans. 2012. ConanDoyle-neg: Annotation of negation in Conan Doyle stories. *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 1563–1568.

Morante, R. y C. Sporleder. 2012. Modality and Negation: An Introduction to the Special Issue. *Computational Linguistics*, 38: 223-260.

Moreno, A., S. López, F. Sánchez, y R. Grishman. 2003. Developing a syntactic annotation scheme and tools for a Spanish treebank. *Treebanks*, 149–163.

Nueva Gramática de la Lengua Española. 2009. Espasa Libros.

Reitan, J., J. Faret, B. Gambäck, B., y L. Bungum. 2015. Negation Scope Detection for Twitter Sentiment Analysis. *Proceedings of the 6thWorkshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 99–108.

Rodriguez-Penagos, C., C. Armentano-Oller, M. Villegas, M. Melero, A. Gonzalez, A., O. de Gibert Bonet, y C. Carrino Pio. 2021. *The Catalan Language CLUB.* arXiv preprint arXiv: 2112.01894.

Seminck, O. (s. f.). *Guide d'Annotation corpus FReND*. https://hackmd.io/@xcEOEFWUR1aRl-Zrtx-lWA/BJav8DMJq#Outil-d%E2%80%99annotation [Recuperado 6 de marxo de 2023].

Tañá, L. 2021. *Anotació del focus de la negació i de la temporalitat en informes mèdics*. [Treball de Fi de Màster, Universitat de Barcelona]. Dipòsit Digital de la Universitat de Barcelona.

Taulé, M., M. Nofre, M. González, y M. A. Martí. 2021. Focus of negation: Its identification in Spanish. *Natural Language Engineering*, 27(2): 131-152.

Tubau, S., y M. T. Espinal. 2010. Doble negació dins l'oració simple en català. *Estudis Romànics*, 34: 145-164.

Wilson, T., J. Wiebe, y P. Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 347–354.

A Anexo 1: Marcadores de negación más frecuentes en las oraciones del corpus

Marcador	Ocurrencias	%
no	3.302	63,72
sense	328	6,33
no… cap	239	4,61
no… res	154	2,97
ni	112	2,16
no… ni	87	1,68
no… pas	69	1,33
tampoc	51	0,98
mai	49	0,95
cap	49	0,95

B Anexo 2: Marcadores de negación más frecuentes en las oraciones procedentes de Datasets

Marcador	Ocurrencias	%
no	302	63,45
sense	40	8,40
no… cap	25	5,25
no… res	19	3,99
no… mai	9	1,89
no… gaire	7	1,47
no… ni	6	1,26
tampoc	6	1,26
no… pas	6	1,26
no… ningú	4	0,84

C Anexo 3: Marcadores de negación más frecuentes en las oraciones procedentes de Foros

Marcador	Ocurrencias	%
no	1.546	63,88
no… res	100	4,13
no… cap	84	3,47
sense	83	3,43
ni	77	3,18
no… ni	43	1,78
mai	33	1,36
tampoc	31	1,28
no… gaire	31	1,28
no… pas	27	1,12

D Anexo 4: Marcadores de negación más frecuentes en las oraciones procedentes de Periódico

Marcador	Ocurrencias	%
no	1.454	63,60
sense	205	8,97
no… cap	130	5,69
no… ni	38	1,66
no… pas	36	1,57
no… res	35	1,53
ni	32	1,40
cap	29	1,27
no… mai	23	1,01
tampoc… no	21	0,92

E Anexo 5: Comparación entre los datos obtenidos en los corpus NoNiRes y NewsCom

	NoNiRes	**NewsCom**
Oraciones	20.541	4.980
Oraciones con alguna estr. de negación	4.488 (21,85 %)	2.247 (45,12 %)
Total estr. de negación	5.182	2.975
Marc. discont.	1.126 (21,73 %)	694 (20,23 %)
Foco elíptico	105 (2,2 %)	161 (5,41 %)