Alumna: Guillermina Pagani

Asignatura: Análisis computacional de datos lingüísticos

Profesor: Javier Vera Zúñiga

**Rastreo y análisis de construcciones multiverbales mediante técnicas computacionales**

1. **Problema de investigación**

Nuestro proyecto de tesis de magíster se centra en la construcción multiverbal monopredicativa *cagarse* + gerundio. La construcción parece codificar diferentes valores semánticos, como puede observarse en los siguientes ejemplos:

1. Me cagué trabajando para comprar mi terreno (=*trabajé mucho*).
2. Se cagó muriendo de un infarto masivo (≠ *se murió mucho*).

Otro contraste entre lo que parecieran ser dos subtipos de esta construcción se relaciona con la posibilidad de negar el gerundio, y evidencia diferencias estructurales además de semánticas:

1. Me cago no saliendo en toda la cuarentena y la vecina se la pasa de joda.
2. \*Se cagó no muriendo[[1]](#footnote-1).

En el siguiente trabajo nos interesa acercarnos al abordaje de este último contraste a partir de un análisis más general de las construcciones multiverbales del español. Por ello, nos preguntamos cómo interactúa la negación con las construcciones multiverbales de verbo no finito.

En la tradición gramatical, el rechazo de la negación independiente para cada verbo es considerado un criterio de diagnóstico de la monopredicatividad de una construcción multiverbal. Así, para Aikhenvald, “the components of a multiverbal construction cannot be negated separately from each other” (2011:19). De manera similar, Haspelmath (2016) defiende la negación única como uno de los pocos criterios fiables para caracterizar las construcciones seriales verbales de distintas lenguas. No obstante, y tal como especifican Jaque *et al* (2021), “la restricción de la negación única podría ser, para el español, un criterio de monoclausalidad demasiado fuerte” (16). En su análisis de la construcción *llegar y* + verbo finito, los autores señalan el hecho de que varias perífrasis del español[[2]](#footnote-2), como las modales (a) y algunas aspectuales de fase (b) (en contraposición a las temporales (c), por ejemplo) admiten negación independiente de ambos verbos. Así:

1. No puede (no) venir / No debe (no) venir.
2. No empieces a (no) venir a clases / No dejó de (no) venir a clases.
3. No va a (\*no) llover. (ejemplos de Jaque *et al* 2021)

El hecho de que una construcción multiverbal admita negación interna resulta relevante ya que esto se relaciona con el nivel de cohesión semántica y de gramaticalización de la construcción. Por otra parte, tiene consecuencias en las clasificaciones propuestas para las perífrasis y en los análisis configuracionales que se realicen del operador negativo. Con este trabajo, esperamos contribuir al debate actual sobre la cuestión (véase también Fernández de Castro 1999, Bravo *et al* 2015, Fábregas & González Rodríguez 2019).

Otra importante motivación de esta investigación es mostrar el valor de las herramientas computacionales y la lingüística de corpus en el estudio de las construcciones multiverbales. Consideramos que la “generalización, confiabilidad y representatividad” de los estudios lingüísticos de corte cualitativo en general, muchas veces puestas en cuestión por “basarse en un pequeño número de textos seleccionados arbitrariamente o posiblemente seleccionados para justificar las nociones preconcebidas del investigador” pueden mejorarse a partir del uso de estas herramientas al “proporcionar una base empírica para validar y reforzar sus hallazgos” (Lee 2019:84). En ese sentido, buscamos realizar un aporte en la elaboración de una metodología computacional para el rastreo y análisis de construcciones.

1. **Objetivos**

Objetivo general:

Caracterizar la relación entre la negación y las construcciones multiverbales de verbo no finito del español.

Objetivos específicos:

1. Rastrear ejemplos de construcciones multiverbales, específicamente perífrasis verbales, en diferentes corpus del español.
2. Determinar con qué verbos auxiliares es admisible la negación interna.
3. Determinar las frecuencias de aparición de la negación con los verbos auxiliares.
4. Proponer una metodología computacional para la búsqueda y análisis de construcciones multiverbales.

En esta etapa del trabajo, nos centraremos en los objetivos específicos 1 y 4, para avanzar lo más posible en su consecución.

1. **Metodología**
   1. **Datos lingüísticos**

Utilizaremos datos de corpus del español disponibles online. Para esta entrega, trabajaremos con el Corpus del Español (Web/Dialects), disponible en www.corpusdelespanol.org, que contiene cerca de 2 billones de palabras de 21 países hispanohablantes en 1.800.000 de textos recogidos de páginas web (60% de blogs) entre 2013 y 2014.

Es importante saber que, dado que algunos de los textos del corpus están protegidos por derechos de autor, en la construcción del corpus se eliminaron 10 de cada 200 palabras y se reemplazaron con una secuencia de caracteres @@@@. Esto significa que los textos contienen el 95% del total de los datos. Según los autores del corpus, la frecuencia relativa de palabras o colocaciones no se ve afectada por estas omisiones dado que ocurren aleatoriamente cada 200 palabras, sin importar el contexto y afectan a todas las palabras por igual.

* + 1. **Descripción de datos de la muestra**

La muestra que analizaremos se extrae de manera *random* del corpus y representa −según lo indicado en la web− un 1/100 del número total de textos. El detalle de composición de la muestra puede observarse en la siguiente tabla:

|  |  |
| --- | --- |
| Lengua de los textos | español |
| Oraciones | 84.373 |
| Tokens | 1.985.509 |
| Types | 88.493 |

* 1. **Técnicas de procesamiento de los datos**

Los archivos .txt extraídos de la web de Corpus del Español han sido codificados en UTF-8 y procesados con módulos de Python 3, como *re* (expresiones regulares) y *csv*. El texto fue sometido a una serie de procedimientos destinados a depurarlo lo más posible. Entre ellos, se destacan los siguientes:

1. Eliminación de encabezados.
2. Eliminación de metadatos con indicaciones y número de texto.
3. Eliminación de puntuación y otros símbolos (@).
4. Eliminación de los espacios en blanco sobrantes.

Hemos decidido no eliminar los caracteres numéricos y los nombres de entidades a fin de evitar que se generaran artificialmente más construcciones (V+V) de las realmente existentes.

El procesamiento y etiquetado de los datos fue realizado con las herramientas Lemmatizer, POS tagging y MorphAnalysis de la librería Spacy (Python 3), que lematizan, asignan categorías gramaticales y brindan información morfológica detallada de cada token del texto. También se han utilizado las herramientas provistas por la librería Pandas.

Con los datos procesados hasta el momento, el corpus tiene un total de 1.985.509 tokens.

1. **Análisis de los datos**

Para el análisis de los datos hemos trabajado con las primeras 1.000 oraciones de la muestra. El total de tokens de estas oraciones es de 24.613, mientras que se registran 5726 types. Hemos utilizado la librería Pandas (Python 3) para crear dataframes con las palabras filtradas.

Inicialmente, hemos filtrado todas las palabras de estas oraciones etiquetadas como verbos (POS=“VERB”), que suman 2.750. Un análisis manual de las primeras 30 oraciones mostró que el filtrado dejaba fuera verbos etiquetados como “ADJ” y “AUX”. En el primer caso, se trataba de participios (como *convocada*), que también resulta necesario incluir para el posterior rastreo de las construcciones multiverbales. Un nuevo filtrado, a partir de la información morfológica (VerbForm= “Part”/participio, “Ger”/gerundio, “Fin”/verbo finito), arroja un total de 4.024 formas verbales. Hemos guardado la posición de estas palabras en el filtrado, para poder buscar las que se encuentren contiguas.

En el análisis llevado a cabo hasta el momento encontramos algunos problemas para tener en cuenta:

1. El etiquetado de los datos no resulta completamente adecuado a nuestros objetivos. Por ejemplo, el token *ver* es etiquetado como sustantivo, con lo que quedan fuera del conteo y filtrado los casos en los que funciona como infinitivo, mientras que *lucha* se etiqueta como verbo en contextos en que funciona como sustantivo.
2. Los pronombres clíticos aparecen pegados al verbo en infinitivo (como *hacerlo*). De todas maneras, esto no genera consecuencias negativas ya que en el análisis de rasgos aparecen los de ambas palabras.
3. Los números del texto y los nombres propios se cuentan como tokens. Como hemos explicado, era necesario conservarlos, pero sería mejor que no formaran parte del recuento total de palabras del texto, para no afectar los posteriores cálculos de frecuencia. Se podría volver a calcular los tokens sin estos elementos una vez extraídas las construcciones.
4. **Conclusiones y perspectivas futuras**

Aunque aún no hemos alcanzado el primer objetivo específico de este trabajo, avanzamos en los procedimientos que nos permitirán hacerlo en la próxima etapa. Se depuró el texto y se agregaron etiquetas de clases de palabra e información morfológica que resulta indispensable para rastrear las construcciones multiverbales. Además, hemos filtrado las formas verbales y guardado su posición, para poder buscar las que se encuentren contiguas. Además, hemos calculado los tokens y types del texto, para poder calcular frecuencias luego. En el proceso se han detectado algunos problemas cuya consideración quedará para las siguientes etapas.

La próxima tarea será filtrar filas cuyo index sea +1 respecto de la fila anterior. Será necesario tomar en cuenta, además, los casos de negaciones interpuestas, cuya búsqueda forma parte de uno de nuestros objetivos. Al mismo tiempo, habrá que decidir si se amplía el criterio de filtrado de modo que sea posible detectar construcciones multiverbales con otros materiales interpuestos, como *también* (en el corpus, p.e., aparece la construcción *estamos también trayendo*, justamente con ese adverbio interpuesto).

Posteriormente, hemos previsto −a partir de la lematización que hicimos de las palabras del texto− hacer una lista de todos los verbos que funcionan como auxiliares en las construcciones detectadas para determinar cuáles de ellos registran casos de negación interna. Por último, calcularemos las frecuencias de aparición de estos verbos con el adverbio de negación.

1. **Bibliografía**

Aikhenvald, A. 2011. Multi-verb constructions: setting the scene. En Aikhenvald, A. y P. Muysken. *Multi-verb constructions. A view from the Americas*. 1-26. Leiden: Brill.

Bravo, A., García Fernández, L. & Krivochen, D. 2015. On Auxiliary Chains: Auxiliaries at the Syntax-Semantics Interface, *Borealis: An International Journal of Hispanic Linguistics 4* (2): 71-101. <https://doi.org/10.7557/1.4.2.3612>

Davies, M. *Corpus del español*. Disponible en <https://www.corpusdelespanol.org/>

Fábregas, A. y González Rodríguez, R. 2019. Perífrasis e inductores negativos: un análisis en términos de dominios, *Onomázein* 43: 95-113. <https://doi.org/10.7764/onomazein.43.10>

Fernández de Castro, F. 1999. *Las perífrasis verbales en el español actual*. Madrid: Gredos.

Haspelmath, M. 2016. The serial verb construction: Comparative concept and cross-linguistic generalizations, *Language and Linguistics* 17/3: 291-319. <https://doi.org/10.1177/2397002215626895>

Jaque, M., M. Covarrubias, C. González, S. Guerrero, F. Hasler y V. Orqueda. 2021. A la altura de las expectativas: interacciones entre la negación y construcciones multiverbales del tipo *llegar y* + VF. En prensa.

Lee, C. 2019. How are ‘immigrant workers’ represented in Korean news reporting? A text mining approach to critical discourse analysis, *Digital Scholarship in the Humanities* 34/1: 82–99, <https://doi.org/10.1093/llc/fqy017>

1. Utilizamos la convención del asterisco (\*) para indicar agramaticalidad, como suele hacerse en la tradición gramatical, específicamente en la teoría generativa. [↑](#footnote-ref-1)
2. Tal como se explica en Jaque *et al* (2018), los casos de perífrasis corresponderían a construcciones multiverbales de verbo infinito (CMVI), tal como la construcción CG. [↑](#footnote-ref-2)