Alumna: Guillermina Pagani

Asignatura: Análisis computacional de datos lingüísticos

Profesor: Javier Vera Zúñiga

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

1. **Problema de investigación**

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. Sin embargo, 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[[1]](#footnote-1), 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)

En el siguiente trabajo nos interesa acercarnos al abordaje de este contraste a partir de la búsqueda de construcciones multiverbales de verbo no finito y negaciones en diversos corpus del español, ya que no existen estudios sobre el tema realizados a partir de corpus reales. Nos preguntamos cómo interactúa la negación con las construcciones multiverbales cuando se observa una gran cantidad de datos: ¿se justifica la postulación del rechazo de la negación interna como criterio de monoclausalidad?, ¿son más las construcciones que la admiten que las que la rechazan?, ¿qué auxiliares encabezan estas construcciones?, ¿hay similitudes entre ellos respecto del valor (TAM) que transmiten o los verbos con los que se combinan?

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 utilizar 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, que, eventualmente, pueda servir para buscar estas construcciones y otras similares en corpus del español o de otras lenguas romances.

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 públicos del español.
2. Determinar con qué verbos auxiliares es admisible la negación interna y calcular frecuencias de aparición.
3. Proponer una metodología computacional para la búsqueda y análisis de las construcciones multiverbales.
4. **Metodología**
   1. **Datos lingüísticos**

Para este trabajo hemos utilizado datos de corpus del español disponibles online: el Corpus del Español (Web/Dialects) y el CORLEC (Corpus Oral de Referencia de la Lengua Española Contemporánea). Del primero es posible descargar una muestra online desde [www.corpusdelespanol.org](http://www.corpusdelespanol.org), mientras que el segundo está disponible para su descarga completa en [www.lllf.uam.es/ESP/Corlec.html](http://www.lllf.uam.es/ESP/Corlec.html). En el caso del CORLEC, hemos trabajado con el repositorio <https://github.com/djeastm/Spanish_Corpus_Analysis_Project>, que ofrece un archivo .txt ya unificado con todos los textos de este corpus.

* + 1. **Corpus utilizados**
       1. **Corpus del Español**

El Corpus del Español (Web/Dialects), disponible en [www.corpusdelespanol.org](http://www.corpusdelespanol.org), 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. La muestra que analizaremos es extraída de manera *random* del corpus[[2]](#footnote-2) y representa -según lo indicado en la web- un 1/100 del número total de textos.

* + - 1. **CORLEC**

El CORLEC se encuentra disponible en [www.lllf.uam.es/ESP/Info%20Corlec.html](http://www.lllf.uam.es/ESP/Info%20Corlec.html). Contiene 1.100.000 palabras de textos grabados y transcritos. Fue elaborado por la Universidad Autónoma de Madrid entre enero de 1991 y febrero de 1992. No se brinda información acerca del origen de los hablantes (variedad lingüística hablada).

Según la información de la web, el corpus incluye los siguientes tipos de texto: administrativos, científicos, conversacionales o familiares, educativos, humanísticos, instructivos, jurídicos, lúdicos, políticos y periodísticos.

* + 1. **Descripción de datos de los corpus**

El detalle de composición de cada corpus puede observarse en la siguiente tabla:

|  |  |  |
| --- | --- | --- |
|  | **Corpus del Español (muestra)** | **CORLEC (completo)** |
| Lengua de los textos | español | español |
| Modalidad de lengua | escrita | oral |
| Oraciones | 84.373 | 69.890 |
| Tokens[[3]](#footnote-3) | 1.985.509 | 946.820 |
| Types | 88.491 | 41.507 |

Tabla 1. Composición de los corpus

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

Los archivos .txt extraídos de la web y del repositorio mencionado han sido codificados en UTF-8 y procesados con módulos de Python 3, como *re* (expresiones regulares) y *csv*. Los textos fueron sometidos a una serie de procedimientos destinados a depurarlos lo más posible. Entre ellos, se destacan los siguientes:

1. Eliminación de encabezados.
2. Eliminación de metadatos con indicaciones y número de texto, hablantes, procedencia, etc.
3. Eliminación de turno de hablante en textos conversacionales.
4. Eliminación de acotaciones sobre gestos, tono de voz, etc.
5. Eliminación de puntuación y otros símbolos (@).
6. 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.

1. **Análisis de los datos**

Para el análisis de los datos hemos trabajado con las primeras 10.000 oraciones de cada corpus. A continuación de detalla el total de tokens y types de estas oraciones para cada corpus (Tabla 2).

|  |  |  |
| --- | --- | --- |
|  | **Corpus del Español** | **CORLEC** |
| Tokens | 236.552 | 105.298 |
| Types | 25.979 | 10.148 |

Tabla 2. Tokens y types en las 10000 oraciones iniciales por corpus

Inicialmente, hemos filtrado todas las palabras de estas oraciones etiquetadas como verbos (POS=“VERB”). La librería Pandas (Python 3) nos permitió crear dataframes con las palabras filtradas. Un análisis manual de las primeras 30 oraciones de cada corpus mostró que el filtrado dejaba fuera verbos etiquetados como “AUX”, por lo que hemos realizado un nuevo filtrado a partir de esta etiqueta para incluir estas formas verbales. Hemos guardado la posición de estas palabras en el filtrado, para poder buscar las contiguas.

En el preprocesamiento y análisis de POS 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; *lucha* se etiqueta como verbo en contextos en que funciona como sustantivo, y *libres* o *lindo*, como verbos cuando aparecen detrás de un auxiliar, lo que en ningún caso es correcto.
2. No se reconocen algunas expresiones lexicalizadas, que funcionan como unidad, como el marcador discursivo “es decir”, que se analiza como AUX + VERB (Inf). Este fue eliminado del listado final de construcciones.
3. Los pronombres clíticos aparecen pegados al verbo en infinitivo (como *hacerlo*). De todas maneras, esto no tiene consecuencias negativas, ya que, en el análisis de rasgos, que es el que hemos considerado para el filtrado, aparecen los de ambas palabras.
4. Los números del texto y los nombres propios se cuentan como tokens. Como hemos explicado, era necesario conservarlos, pero a efectos de posibles cálculos de frecuencia, debe aclararse que formaron parte del recuento total de palabras del texto.

Posteriormente, hemos realizado el filtrado de las construcciones verbales (auxiliares más verbos en infinitivo, participio y gerundio, VerbForm= “Inf”, “Part” y “Ger”) más las negaciones (ADV con el rasgo “Polarity Neg”) que aparecían en posiciones contiguas a alguna forma verbal. A continuación, se realizó una revisión manual de las primeras 100 construcciones de cada lista, para ajustar el filtrado. En base a esta revisión, se aplicaron los siguientes procesos:

1. Se recategorizaron verbos como *seguir* y *soler*, que funcionan como auxiliares en perífrasis verbales documentadas (García Fernández 2006). Se los re-etiquetó como AUX. De esta manera, se incluyeron construcciones que quedaban fuera del primer filtro, como *siga siendo* (progresiva) o *suele empobrecer* (frecuentativa o habitual).
2. Se eliminaron las construcciones que solo constaban de una negación más una forma verbal.
3. Se eliminaron las construcciones que solo constaban de negaciones.
4. Se eliminaron las negaciones que ocupaban la primera y última posición de cada construcción, para conservar solo los casos de negación interna.
5. Se realizó un nuevo filtro para conservar solamente las construcciones que tuvieran un auxiliar en la primera posición.
6. Se eliminaron las combinaciones del verbo *ser* + infinitivo, que no constituyen una construcción multiverbal (esto dejó fuera expresiones como *es decir*), a diferencia de *ser* + participio.
7. A medida que se realizaron los procesos anteriores, se fueron eliminando las construcciones que quedaron con un solo constituyente.

Un contraste entre la lista original de construcciones y la lista ajustada mostró que los cambios fueron adecuados para eliminar secuencias que no correspondían a construcciones multiverbales reales. En el caso específico del CORLEC, fue necesario aplicar recursivamente algunos filtrados (como la eliminación de las negaciones finales o la de las construcciones constituidas por una o ninguna palabra (strings vacíos)) dada la gran cantidad de repeticiones que se registraban en los enunciados (p.e. *no no es*), presumiblemente por la naturaleza oral del corpus.

Las siguientes tablas muestran el total de construcciones halladas en cada corpus (Tabla 3) y un detalle de la composición de estas construcciones (Tabla 4):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Corpus del Español** | | **CORLEC** | |
|  | **Tokens** | **Types** | **Tokens** | **Types** |
| Construcciones multiverbales de verbo no finito (CMVNF) | 3416 | 1767 | 2070 | 836 |

Tabla 3. Cantidad de CMVNF por corpus

|  |  |  |
| --- | --- | --- |
| *Tokens* | **Corpus del Español** | **CORLEC** |
| Formas verbales | 20306 | 11819 |
| Verbos auxiliares que forman parte de las CMVNF[[4]](#footnote-4) | 3933 | 2488 |
| Construcciones de gerundio | 509 | 357 |
| Construcciones de infinitivo | 1751 | 528 |
| Construcciones de participio | 1606 | 1173 |
| CMVNF con negación interna | 18 | 26 |

Tabla 4. Composición de las CMVNF por corpus

En relación con los resultados, debe aclararse que hay errores por inclusión y por exclusión. La búsqueda ha mostrado limitaciones por la misma ambigüedad existente en el lenguaje y hemos tenido que tomar ciertas decisiones. Así, las palabras han recibido una etiqueta POS a partir de la cual se han realizado los filtrados, pero, en muchos casos, podrían recibir otra. Para dar un ejemplo, una secuencia como *es reducida* puede etiquetarse como AUX + participio o como AUX + adjetivo, pero la elección por uno u otro análisis solo podría hacerse sin fallos si contáramos con un contexto oracional que permitiera la desambiguación (en *El presupuesto es reducido por el Congreso por segunda vez consecutiva en el año*, en donde tenemos una construcción multiverbal pasiva, o *El presupuesto con el que cuenta para hacerlo es reducido*, donde hay una lectura adjetiva predominante del segundo componente de la secuencia)[[5]](#footnote-5). Hemos optado por dejar fuera las palabras etiquetadas como adjetivos por el POS tagger de SpaCy porque su inclusión en un primer momento de prueba llevó a una cantidad de errores de detección mucho mayor que la cantidad de potenciales CMVNF excluidas. Otro caso a tomar en cuenta fue el de los de verbos que pueden funcionar como AUX o verbos plenos, como *saber* en *saber prevenir*, que hemos optado por incluir dentro del filtrado de CMVNF, pero que podrían analizarse como verbo principal más oración subordinada de infinitivo (como CMVNF, *saber* + infinitivo tiene un valor frecuentativo en donde *saber* equivale a *soler*, mientras que en el análisis alternativo no hay una única construcción sino dos verbos pertenecientes a cláusulas distintas: *saber* y *prevenir*: *Hizo un curso y sabe prevenir incendios*)[[6]](#footnote-6). Sabemos que en ambos casos es posible elegir otra opción (que también llevará a errores), pero lo especificamos para mostrar la complejidad del proceso de rastreo de construcciones y explicitar las alternativas tomadas en el diseño de la metodología.

Por último, respecto de los resultados, y como puede desprenderse de la lectura de la Tabla 4, hemos encontrado muy pocos casos de CMVNF con negación interna. En la gran mayoría de los casos, se deben a errores del filtrado o a repeticiones y correcciones. Por ejemplo: en *puede ocurrirte no puede ocurrir* (CORLEC) hay, evidentemente construcciones multiverbales, pero el adverbio de negación funciona corrigiendo la primera y no es una negación interpuesta. En numerosas ocasiones, ni siquiera hay CMVNF, como en *hay no es* (CORLEC), en donde la secuencia se compone, efectivamente de dos auxiliares y una negación (criterio de inclusión para las CMVNF), pero claramente no hay una construcción. Una revisión manual de los 44 casos detectados mostró que solo 3 son realmente casos de CMVNF con negación interna, los 3 con el verbo *poder*: *podría no volver* (60393-69395, CESP) y *podría no estar* (75235-75237 y 75262-75264, CORLEC).

1. **Conclusiones y perspectivas futuras**

A lo largo de este trabajo, hemos avanzado en la construcción de una metodología para la búsqueda de construcciones multiverbales: se depuraron los textos de los corpus y se agregaron etiquetas de clases de palabra e información morfológica que resultaban indispensables para el rastreo. Se filtraron las formas verbales y se guardó su posición, para buscar las contiguas. A partir de esta información y del establecimiento (y revisión) de criterios de inclusión y exclusión, elaboramos un corpus de construcciones multiverbales de verbo no finito, disponible para todo público en el repositorio Github: <https://github.com/paganiguillermina/Corpus-de-construcciones-multiverbales-del-espanol>. En el proceso se han detectado algunos problemas que intentamos eliminar, aunque aún se registran errores, mayormente, por inclusión.

Hemos buscado, asimismo, los casos de negaciones interpuestas en estas CMVNF, cuya búsqueda formaba parte de uno de nuestros objetivos. Aunque los resultados fueron muy limitados en este punto, y no nos han permitido realizar cálculos de frecuencia como esperábamos, creemos que pueden sacarse algunas conclusiones interesantes respecto de la relación entre CMVNF y negación. El análisis computacional detectó más construcciones con negación de las reales, pero el número sigue siendo muy reducido (44) respecto del total de construcciones (5486). Esto es incluso cierto para el corpus oral que hemos analizado (CORLEC), lo que resulta interesante si se toma en cuenta que en la oralidad la sintaxis suele ser más flexible que en la escritura. Que los casos de CMVNF con negación interna sean tan limitados -aun con sobredetección- pareciera mostrar la poca frecuencia de este fenómeno y podría ser un indicio de que el criterio de rechazo de la negación independiente para cada verbo es un criterio fuerte para el diagnóstico de la monopredicatividad de una construcción multiverbal.

Por otro lado, la realización de este trabajo ha puesto en evidencia la dificultad de trabajar con el nivel sintáctico en corpus orales. El rastreo y análisis computacional de construcciones en el caso del CORLEC ha requerido aplicar ciclos recursivos de eliminación de datos por la gran cantidad de repeticiones, correcciones y fallos cometidos por los hablantes (p.e. *era no no*, *puedo arreglar no no*), e incluso así se registran más errores en la lista final de CMVNF de este corpus que en el CESP (escrito). El análisis computacional de construcciones pareciera tener más limitaciones con corpus orales.

Más allá de esta salvedad, consideramos que este trabajo ha mostrado el valor de la lingüística computacional en el análisis de datos lingüísticos, y su relevancia en el objetivo de fortalecer la “generalización, confiabilidad y representatividad” de los estudios lingüísticos de corte cualitativo. Un análisis preliminar y muy sucinto de más de 340.000 palabras mostró la escasa aparición de la negación interna en el contexto de una CMVNF, sustentando más empíricamente criterios establecidos en la bibliografía. Sería necesario investigar qué sucede en otras lenguas para otorgar mayor solidez a estas consideraciones, lo que queda para trabajos futuros.

Por último, creemos que otra línea de análisis interesante y complementaria sería la que procure trabajar computacionalmente con corpus ya analizados sintácticamente (p.e. Universal Dependencies), lo que podría reducir el número de errores en la búsqueda de ciertas construcciones.

1. **Corpus utilizados**

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

Laboratorio de Lingüística Informática, Universidad Autónoma de Madrid. *Corpus Oral de Referencia de la Lengua Española Contemporánea* (CORLEC). Disponible en <http://www.lllf.uam.es/ESP/Info%20Corlec.html>

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>

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.

García Fernández, L. (dir.). 2006. *Diccionario de perífrasis verbales*. 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>

Yllera, A. 1999. Las perífrasis verbales de gerundio y participio. En Bosque, I. y V. Demonte (eds.). *Gramática descriptiva de la lengua española*. 3391-3441. Madrid: Espasa Calpe.

1. **Librerías y repositorios utilizados**

Djeastm. Repositorio Github *Spanish\_Corpus\_Analysis\_Project*. Disponible en <https://github.com/djeastm/Spanish_Corpus_Analysis_Project>

Pandas, versión 1.3.4, disponible en <https://pandas.pydata.org/>

SpaCy, versión 3.2, disponible en <https://spacy.io/>

1. 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-1)
2. Lamentablemente, es la única muestra que se ofrece en la web de manera gratuita. No es posible descargar un paquete de datos diferente aunque se intente otra vez. [↑](#footnote-ref-2)
3. Los corpus suelen medirse en cantidad de palabras o tokens. Consideramos *token* como la ocurrencia individual de una cadena de caracteres separada por espacio en un texto y *type* como la unidad abstracta que representa esas expresiones lingüísticas concretas, es decir, las diferentes cadenas en un texto. [↑](#footnote-ref-3)
4. Estos superan en número a la cantidad de CMVNF porque una misma construcción puede construirse con uno o más auxiliares (p.e. *deben* AUX *ser* AUX *analizados*, la primera de la lista en el CESP) [↑](#footnote-ref-4)
5. El mismo contexto oracional podría, incluso, no desambiguar la clase de cada palabra: *El presupuesto es reducido*. [↑](#footnote-ref-5)
6. La combinación de verbo flexionado más forma no finita “puede funcionar como un único núcleo verbal o como dos núcleos verbales”, por lo que “fuera de todo contexto, una misma oración puede resultar ambigua, es decir, una misma secuencia puede responder a dos significados léxicos y gramaticales” (Yllera: 3395). [↑](#footnote-ref-6)