ADAPTACIÓN AUTOMÁTICA A LECTURA FÁCIL DE PERÍFRASIS VERBALES EN ESPAÑOL[1]

AN AUTOMATIC EASY-TO-READ ADAPTATION OF VERBAL PERIPHRASES IN SPANISH

Isam Diab-Lozano
Universidad Politécnica de Madrid

Mari Carmen Suárez-Figueroa
Universidad Politécnica de Madrid

Alejandro Muñoz-Navarro
Real Academia Española

RESUMEN
Las perífrasis verbales son estructuras lingüísticas complejas para distintos grupos de población durante el proceso de comprensión lectora. La Metodología de Lectura (LF) surgió para facilitar la comprensión lectora, mediante una serie de pautas y recomendaciones para presentar documentos claros y fáciles de entender. Una de estas pautas se centra en evitar el uso de perífrasis verbales, por su complejidad para personas con dificultades de comprensión. Para abordar este desafío, hemos desarrollado un método inicial de adaptación automática de perífrasis verbales en español, basado en (a) reglas y patrones, parte de la llamada Inteligencia Artificial (IA) simbólica, y (b) las pautas de LF. Hemos implementado el método como un servicio web, cuyos resultados han sido evaluados por personas con discapacidad cognitiva. Además, hemos usado grandes modelos de lenguaje (LLM), parte de la IA subsimbólica, para adaptar automáticamente un tipo específico de perífrasis verbales que nuestro método no era capaz de adaptar.

PALABRAS CLAVE: Lectura Fácil, Accesibilidad Cognitiva, Traducción automática, Adaptación de textos

ABSTRACT
Verbal periphrases are considered complex linguistic structures for different groups of the population during the reading comprehension process. The Easy-to-Read (E2R) Methodology was created to facilitate reading comprehension. This

[1] Esta investigación se enmarca en el proyecto 'AILF – Inteligencia Artificial para la Lectura Fácil' (Ref. APOYO-JOVENES-21-8AV1UF-119-OEUU22) y ha sido financiada también por la ayuda predoctoral PIPF-2022/COM-25762 de la Comunidad de Madrid. Agradecemos la colaboración de Isa Cano, María José Sánchez, ACCEDES y sus validadores con discapacidad intelectual, así como del COFOIL 'cuarentainueve' (AMP), y a Arminda Moreno por su apoyo en el análisis de los datos.

CC BY NC ND

methodology includes a set of guidelines and recommendations to present clear and easily understood documents. One of these guidelines relies on avoiding the use of verbal periphrases, due to their complexity for individuals with reading difficulties. To aid comprehension, we have developed an initial method for automatically adapting verbal periphrases in Spanish. This method is based on (a) rules and patterns, part of the so-called symbolic Artificial Intelligence (AI), and (b) the E2R guidelines. The method has been implemented as a web service, whose results have been evaluated by people with cognitive disabilities. In addition, we explore the use of Large Language Models (LLMs), part of the sub-symbolic AI, to automatically adapt a specific type of verbal periphrases overlooked by our method.

KEYWORDS: Easy-to-Read, Cognitive Accessibility, Machine-Translation, Text Adaptation

1. INTRODUCCIÓN

Las perífrasis verbales son construcciones gramaticales formadas por dos o más verbos que, en conjunto, funcionan como una sola unidad verbal (García González, 1992; Torrego, 1988). Aunque son frecuentes en el discurso cotidiano, estas estructuras pueden dificultar la comprensión lectora, especialmente entre personas con dificultades cognitivas o estudiantes de español como lengua extranjera (Gil, 2011). Por este motivo, la metodología de lectura fácil (LF o *Easy-to-Read (E2R)* en inglés) recomienda evitarlas en los textos adaptados (AENOR, 2018; Inclusion Europe, 2009; Nomura et al., 2010).

El propósito de la metodología LF es garantizar un acceso equitativo a la información para colectivos con necesidades específicas de comprensión, como personas con discapacidad intelectual, lectores con bajo nivel de alfabetización o dominio limitado del idioma. Esta metodología ofrece recomendaciones tanto lingüísticas – como emplear estructuras sintácticas sencillas y evitar construcciones complejas – como de presentación – como dividir las ideas en párrafos o incluir apoyos visuales (AENOR, 2018).

Actualmente, la adaptación de textos a lectura fácil se realiza de forma manual, a través de un proceso iterativo que incluye: (1) análisis del cumplimiento de las directrices LF en el documento original; (2) transformación del contenido por parte de especialistas; y (3) validación por un equipo profesional que incluye personas con discapacidad cognitiva, conocidas como validadoras (AENOR, 2018). Este proceso, aunque riguroso, resulta costoso y demanda mucho tiempo. De ahí surge la necesidad de desarrollar herramientas tecnológicas que puedan apoyar o automatizar parte de esta tarea.

En este contexto, nuestra línea de investigación se sitúa en la intersección entre accesibilidad y tecnologías lingüísticas, con el objetivo de automatizar el análisis y la adaptación de textos en español según los principios de la LF. Buscamos, así, producir versiones comprensibles que mantengan el significado del texto original.
En particular, este trabajo se centra en una de las directrices más desatendidas en la práctica: la necesidad de evitar las perífrasis verbales (excepto aquellas formadas con verbos modales como “deber”, “querer”, “saber” o “poder”). Estas estructuras, al combinar múltiples formas verbales, añaden complejidad innecesaria a las oraciones y dificultan la comprensión (Bott et al., 2012).

CC BY NC ND

Desde una perspectiva cognitiva, la comprensión lectora implica múltiples variables relacionadas con el lector: desde características personales hasta competencias lingüísticas y habilidades lectoras (Mata et al., 2007). Sin embargo, no todas las personas disponen de estos recursos en igual medida, lo que justifica el uso de metodologías como la LF para facilitar el acceso a la lectura, la cultura y la información.

En cuanto a los desarrollos tecnológicos aplicados a la LF en español, aún son escasos. Se han propuesto herramientas como Easy-to-Read Advisor (Suárez-Figueroa et al., 2020) y su mejora, FACILE (Suárez-Figueroa et al., 2022); Comp4Text (Iglesias et al., 2020), que evalúa la legibilidad de textos web; y E2R-Helper (Díez et al., 2019), un asistente de lectura fácil. En el ámbito de la transformación automática de textos destacan sistemas como Simplext (Saggion et al., 2015), LexSiS (Bott et al., 2012), DysWebxia (Rello et al., 2013) o EASIER (Moreno et al., 2020). Sin embargo, ninguno de estos trabajos aborda de forma específica la identificación y adaptación de perífrasis verbales a formas más accesibles.

Con la intención de cubrir este vacío y contribuir a mejorar la comprensión lectora en poblaciones vulnerables, este estudio se propone identificar automáticamente determinadas perífrasis verbales y transformarlas en estructuras más sencillas y comprensibles. Para ello, hemos desarrollado un método declarativo[2] (Russell y Norvig, 2021) que permite adaptar aquellas perífrasis que no cumplen las directrices de la LF. Además, hemos implementado un prototipo funcional basado en dicho método, que puede considerarse una suerte de traducción intralingüística automatizada, orientada a la accesibilidad, y lo hemos validado con personas con discapacidad cognitiva. En los casos en los que el sistema no logra generar una adaptación adecuada, exploramos también el uso de modelos de lenguaje de gran escala (en inglés, *Large Language Models*, LLM) como alternativa basada en técnicas de IA subsimbólica.

No obstante, conviene señalar desde el inicio que este estudio adopta un carácter exploratorio. La ausencia de corpus anotados en español para perífrasis verbales limita tanto el alcance del método simbólico como la posibilidad de entrenar modelos neuronales más robustos. Del mismo modo, los experimentos con usuarios finales se basan en una muestra reducida y geográficamente concentrada, lo que condiciona la generalización de los resultados. Estas limitaciones forman parte inherente del estado actual del campo, pero justifican la necesidad de iniciar trabajos como el que aquí se presenta.

El resto del artículo se organiza de la siguiente manera: en la Sección 2 se describen las características lingüísticas de las perífrasis verbales y los enfoques automáticos existentes para su identificación y adaptación. En la Sección 3 se presenta nuestra propuesta para adaptar perífrasis verbales en textos en español a lectura fácil, incluyendo el método, la prueba de concepto y una validación con personas con discapacidad cognitiva. Además, se introduce un enfoque preliminar

[2] El conocimiento humano se representa de forma explícita mediante estructuras declarativas (por ejemplo, hechos y reglas). Esta forma de proceder forma parte de la denominada inteligencia artificial simbólica.

CC BY NC ND

que emplea LLM[3] para adaptar un tipo específico de perífrasis verbales. Finalmente, la Sección 4 recoge las conclusiones y futuras líneas de trabajo.

2. ESTADO DE LA CUESTIÓN

Esta sección se centra en dos aspectos fundamentales para contextualizar nuestro trabajo. En primer lugar, se presentan algunas consideraciones lingüísticas esenciales sobre las perífrasis verbales. En segundo lugar, se ofrece una revisión de los enfoques automáticos orientados a la identificación y/o transformación de estas estructuras en formas verbales más simples.

2.1. Aproximación lingüística a las perífrasis verbales

Desde una perspectiva lingüística, una perífrasis verbal es una construcción formada por dos verbos que constituyen un único núcleo verbal (García González, 1992; Torrego, 1988). Por ejemplo, en la oración "Ana y su madre solían ir a la playa en verano", la combinación de los verbos "solían" e "ir" puede entenderse como equivalente a la forma simple "iban".

Las perífrasis verbales son combinaciones sintácticas en las que un verbo auxiliar modifica a un verbo principal, normalmente en forma no personal (infinitivo, gerundio o participio), sin que se generen dos proposiciones independientes. En determinadas construcciones, como "empezar a comer", pueden aparecer elementos gramaticales intermedios, como preposiciones, entre el auxiliar y el verbo principal. Además, los verbos auxiliares pueden encadenarse, lo que da lugar a perífrasis múltiples, como en "Álex debe de ir a empezar a cantar".

En cuanto a las propiedades de ambos verbos, el verbo principal mantiene sus complementos y características gramaticales (por ejemplo, "Ana va a comer pescado en casa de su madre"), mientras que el verbo auxiliar carece de estructura argumental propia y no aporta complementos a la construcción (RAE y ASALE, 2009).

La clasificación de las perífrasis verbales se basa en criterios sintácticos y semánticos. Desde el punto de vista sintáctico, se distinguen según la forma no personal del verbo principal: perífrasis de infinitivo, de gerundio y de participio. Semánticamente, se diferencian entre perífrasis modales y tempo-aspectuales (RAE y ASALE, 2009). Las modales expresan la actitud del hablante hacia la acción (capacidad, posibilidad, obligación, etc.), mientras que las tempo-aspectuales indican cómo se concibe el desarrollo de la acción, diferenciando, por ejemplo, entre perífrasis de fase (incoativas, terminativas, iterativas) y escalares. Ejemplos de estas últimas son "Juan dejó de fumar hace un año" (perífrasis terminativa) o "María terminó leyéndose el libro" (perífrasis escalar). Ambos criterios, sintáctico y semántico, se combinan, de manera que dentro de cada tipo semántico es posible identificar los tres tipos sintácticos.

En el contexto de la metodología de lectura fácil, las perífrasis verbales se consideran construcciones que deben evitarse por su complejidad estructural y su

[3] Durante el resto del artículo haremos uso de esta sigla, pues en español no existe una sigla estandarizada para este término.

CC BY NC ND

impacto negativo en la comprensión lectora (AENOR, 2018; García Muñoz, 2012; Celayeta Gil, 2011).

La selección de los tipos de perífrasis tratados en este trabajo responde a criterios lingüísticos vinculados directamente con la comprensión lectora. Las perífrasis tempo-aspectuales y temporales se priorizan porque codifican valores aspectuales –como inicio, continuidad, interrupción o frecuencia de la acción – que no se derivan de forma transparente de la morfología verbal y exigen inferencias adicionales por parte del lector, lo que incrementa la carga cognitiva. En cambio, otras perífrasis más frecuentes y semánticamente más transparentes, como las modales básicas ("deber", "querer", "saber" o "poder"), se consideran admisibles en el marco de la lectura fácil y no requieren una intervención sistemática. La adaptación de las perífrasis seleccionadas tiene, además, implicaciones lingüísticas relevantes: al reformular una construcción perifrástica en una forma verbal simple o acompañada de un adverbio temporal, se modifica la manera en que se expresa el aspecto verbal y, con ello, la representación temporal del evento en el discurso. Esto es especialmente evidente en las perífrasis terminativas o incoativas, cuyo valor aspectual depende en gran medida del contexto narrativo. Por tanto, la intervención automática no se limita a la sustitución formal, sino que implica decisiones semánticas orientadas a equilibrar la simplificación estructural con la preservación del significado.

2.2. Enfoques automáticos para el tratamiento de las perífrasis verbales

En el ámbito hispánico, los trabajos orientados a la identificación o adaptación automática de perífrasis verbales son escasos. Una excepción es la propuesta de Castro-Sánchez et al. (2015), quienes diseñan un método semiautomático para la detección de colocaciones en definiciones verbales del diccionario de la RAE. Algunas de estas definiciones incluyen perífrasis verbales, especialmente modales, que pueden identificarse mediante técnicas de procesamiento del lenguaje natural.

Otro trabajo destacable es el de Aparicio Mera (2015), centrado en la representación computacional de las perífrasis de fase. En esta investigación se analizan las restricciones combinatorias de estas estructuras y se propone una modelización basada en gramáticas formales (a nivel sintáctico) y reglas heurísticas (a nivel semántico).

Aunque no se centran exclusivamente en las perífrasis, algunos estudios sobre simplificación léxica han abordado la sustitución de determinados verbos por formas más simples, como en el caso de Drndarevic et al. (2012), en continuidad con el trabajo previo de LexSIS (Bott et al., 2012).

Fuera del ámbito hispánico, en lengua alemana, Holz y Biemann (2008) han explorado la detección de perífrasis equivalentes a compuestos nominales a partir de corpus, mediante un enfoque no supervisado y sin conocimiento previo.

No obstante, ninguno de los trabajos mencionados aborda específicamente la identificación ni la adaptación[4] de perífrasis verbales en español a formas más simples cercanas a la lectura fácil.

[4] La adaptación de textos siempre tiene como objetivo transformar los textos para satisfacer las necesidades de una audiencia específica, mientras que la simplificación textual tiende a reducir la complejidad de los textos y no siempre toma en cuenta al usuario final (Saggion, 2022).

CC BY NC ND

3. PROPUESTA DE ADAPTACIÓN DE PERÍFRASIS VERBALES EN ESPAÑOL

Como se indicó en la sección introductoria, este trabajo se centra en la identificación automática y la adaptación a lectura fácil de las perífrasis verbales. En esta sección, presentamos nuestro método declarativo para la identificación y adaptación de aquellas perífrasis que no cumplen con los principios de la metodología de lectura fácil. Asimismo, describimos la prueba de concepto desarrollada a partir de este método y su evaluación, y exponemos una primera aproximación al uso de grandes modelos de lenguaje (LLM) como apoyo en aquellos casos en que el método no ofrece un rendimiento satisfactorio.

3.1. Método para la adaptación de perífrasis verbales a lectura fácil

El método propuesto para adaptar automáticamente las perífrasis verbales en textos en español, según los principios de la metodología de lectura fácil, se compone de tres actividades principales: (1) Procesamiento del Lenguaje Natural (PLN), que incluye tareas de toquenización, etiquetado morfosintáctico, análisis morfológico y detección de dependencias; (2) Identificación de perífrasis verbales; y (3) Adaptación de las perífrasis verbales.

Las actividades 2 y 3 se fundamentan en un enfoque basado en reglas, empleando técnicas de reconocimiento de patrones. En el ámbito de la inteligencia artificial, una regla se define como una estructura del tipo "SI-condición-ENTONCES-acción", en la que la condición se expresa a través de un patrón lingüístico. Para este trabajo, se diseñaron patrones sintácticos inspirados en las conocidas estructuras de Hearst (1992).

El enfoque empleado se enmarca en la denominada inteligencia artificial simbólica (Russell y Norvig, 2021). Optamos por este paradigma simbólico para evaluar de forma preliminar la viabilidad de identificar y adaptar perífrasis verbales al formato de lectura fácil, ya que: (a) actualmente no existe un corpus amplio de textos en español anotado con información sobre perífrasis verbales que permita aplicar métodos subsimbólicos, y (b) la creación de dicho corpus es una tarea compleja y de gran exigencia temporal.

A continuación se explican en detalle las actividades en las que se basa el método desarrollado.

Actividad 1: Procesamiento del Lenguaje Natural

En la tarea de etiquetado morfosintáctico (*Part-of-Speech (PoS)* en inglés) utilizamos las etiquetas universales[5] ROOT para el verbo principal y AUX para los verbos auxiliares. Además, consideramos el atributo lemma, que proporciona la forma en infinitivo del verbo, y el atributo morph, que recoge la información sobre el modo, tiempo y persona verbal.

[5] Pueden consultarse en línea en el siguiente enlace: https://universaldependencies.org/u/pos/

CC BY NC ND

Actividad 2: Identificación de perífrasis verbales

Para identificar las perífrasis verbales en oraciones en español, hemos diseñado los siguientes patrones de identificación (PI):

PI1 ::= AUX SCONJ[lemma="que"] VERB[VerbForm=Inf]
PI2 ::= AUX ADP VERB[VerbForm=Inf]
PI3 ::= AUX VERB[VerbForm=Inf]
PI4 ::= AUX VERB[VerbForm=Ger]
PI5 ::= AUX VERB[VerbForm=Part]

Donde SCONJ representa una conjunción subordinante (i.e. "que"), ADP una preposición, AUX el verbo auxiliar y VERB el verbo principal de la oración, que puede aparecer en forma de infinitivo (VerbForm=Inf), gerundio (VerbForm=Ger) o participio (VerbForm=Part).

Actividad 3: Adaptación de perífrasis verbales

La adaptación de las perífrasis identificadas se basa en un conjunto de patrones de adaptación (PA) cuyo objetivo es generar una versión comprensible conforme a los principios de la lectura fácil. Los patrones de adaptación dependen del tipo de perífrasis detectada.

En el caso de las perífrasis modales, reconocidas mediante los patrones PI1 y PI3, la adaptación solo se aplica cuando se encuentra la estructura "tener que + infinitivo", seguida de un verbo en participio. Para estos casos, se aplica el siguiente patrón de adaptación:

PA1 ::= VERB[VerbForm=Fin, Tense=Fut] VERB[VerbForm=Part]

La conjugación en futuro del verbo en infinitivo se genera a partir de su lema y la información morfológica extraída en la actividad 1. Por ejemplo, la oración original "Tiene que haber sido el viento" se transforma en "Habrá sido el viento".

Para las perífrasis tempo-aspectuales con infinitivo (PI2) y con gerundio (PI4), se emplea el siguiente patrón:

PA2 ::= VERB[VerbForm=Fin, <AuxMorph>]

Aquí el verbo principal aparece originalmente en infinitivo (VerbForm=Inf) o gerundio (VerbForm=Ger), y la forma finita generada hereda los rasgos morfológicos del auxiliar (persona, número, tiempo y modo). Por ejemplo, "Se meterá a jugar al baloncesto" se adapta como "Jugará al baloncesto".

En las perífrasis tempo-aspectuales con participio (PI5), se adapta únicamente si el verbo auxiliar no es "ser" ni "estar". En este caso se utiliza el patrón:

PA3 ::= VERB[VerbForm=Fin, <AuxMorph>]

Dentro de las perífrasis tempo-aspectuales, se identificó una subcategoría relevante: las perífrasis temporales, frecuentes en textos en español. Por ello, también se contemplan en el método.

CC BY NC ND

Para las perífrasis temporales con infinitivo (VerbForm=Inf) (PI2) con el auxiliar "soler", se aplica el patrón:

PA4 ::= VERB[VerbForm=Fin, <AuxMorph>] ADV

Donde la marca temporal (ADV, como etiqueta genérica) puede ser "con frecuencia" o "a menudo". Por ejemplo, "Ana suele comprar los domingos" se adapta como "Ana compra con frecuencia los domingos".

Cuando la perífrasis temporal con infinitivo es detectada mediante PI3 y el auxiliar es "volver", se emplea:

PA5 ::= VERB[VerbForm=Fin, <AuxMorph>] ADV

Con marcas temporales como "nuevamente" o "de nuevo".

Para las perífrasis temporales con gerundio (VerbForm=Ger) (PI4), si el auxiliar es "seguir", se utiliza:

PA6 ::= VERB[VerbForm=Fin, <AuxMorph>] ADV

Y si el auxiliar es "continuar", el patrón es:

PA7 ::= VERB[VerbForm=Fin, <AuxMorph>] ADV

En ambos casos, las marcas temporales son "todavía" o "aún".

A continuación, se presenta la Tabla 1 con ejemplos de oraciones originales que contienen distintos tipos de perífrasis verbales y sus correspondientes versiones adaptadas a lectura fácil.

Tabla 1. Ejemplos de adaptación de perífrasis para cada tipología

Tipo de perífrasis	Oración original	Oración adaptada
Modal (infinitivo)	Tiene que haber sido el viento.	Habrá sido el viento.
Aspecto-temporal (infinitivo)	Estoy por conocer a tu padre.	Conozco a tu padre.
Participio	Pedro nos tiene dicho que no.	Pedro nos dice que no.
Gerundio	Continuaron comiendo después de hablar.	Comieron aún después de hablar.
Temporal	Ana suele comprar los domingos.	Ana compra con frecuencia los domingos.
Múltiple	Pablo se echó a llorar al ver los lugares donde solían veranear juntos.	Pablo lloró al ver los lugares donde veraneaban con frecuencia juntos.

3.2. Prueba de concepto para la adaptación de perífrasis verbales a lectura fácil

Hemos desarrollado una prueba de concepto, implementada como un servicio web RESTful[6], orientada a la identificación y adaptación de perífrasis tempo-aspectuales

[6] El servicio no se encuentra actualmente disponible por razones de privacidad.

CC BY NC ND

en textos en español. Dicha prueba de concepto está basada en el método descrito en la sección anterior.

Este servicio recibe como entrada un texto redactado en español y ofrece como salida una adaptación a lectura fácil de las perífrasis verbales presentes en dicho texto. Tanto la entrada como la salida del servicio están estructuradas en objetos JSON[7]. Además, el servicio almacena internamente las perífrasis identificadas en formato JSON.

La prueba de concepto ha sido desarrollada en el lenguaje de programación Python, en su versión 3.9. El servicio emplea la biblioteca spaCy[8] para tareas basadas en PLN como la segmentación de oraciones, la toquenización y el análisis lingüístico necesario, que incluye la extracción de atributos morfológicos y sintácticos, así como el etiquetado de categorías gramaticales. Asimismo, se ha integrado la biblioteca mlconjug[9] para realizar conjugaciones verbales precisas.
Debido a ciertas imprecisiones ocasionales en el etiquetado de lemas y en la detección de detalles morfológicos de los verbos nominales por parte de spaCy, hemos incorporado la biblioteca SpellChecker[10] para corregir estos casos. Por ejemplo, cuando se encuentra un verbo como "ponerse", la prueba de concepto se asegura de identificar correctamente el lema "poner" y de añadir la información correspondiente al uso reflexivo, evitando así confiar en el lema incorrecto "ponerser", que spaCy puede generar y que SpellChecker identifica acertadamente como inexistente. En estos casos, es necesario ajustar el lema para facilitar la adaptación.

Para verificar los resultados del servicio, se elaboró manualmente un conjunto de pruebas unitarias. Estas pruebas incluyen el texto original, las perífrasis identificadas y las adaptaciones esperadas. Las oraciones originales fueron extraídas de fuentes en línea[11] que contienen ejemplos y clasificaciones de perífrasis verbales; por su parte, las adaptaciones esperadas fueron elaboradas manualmente. Estas adaptaciones se compararon con las versiones generadas automáticamente por la prueba de concepto. Dado el esfuerzo requerido para la recopilación de oraciones, el conjunto de pruebas es limitado e incluye 87 oraciones en español[12]. Para llevar a cabo estas pruebas, se desarrolló un *script* personalizado en Python que compara la salida del servicio con las anotaciones incluidas en las pruebas. De los casos evaluados, 76 fueron considerados correctos, lo que representa un 87 % de acierto.

Además de esta verificación durante el desarrollo del servicio, se utilizó la prueba de concepto, basada en el método declarativo descrito, con un conjunto de 64

[7] Un JSON (JavaScript Object Notation) es un formato ligero de intercambio de datos, fácil de leer y escribir para las personas, y fácil de interpretar y generar por las máquinas. Se utiliza ampliamente para enviar datos entre un servidor y una aplicación web.
[8] La documentación oficial de esta biblioteca está disponible en: https://spacy.io/
[9] La documentación oficial de esta biblioteca está disponible en: https://mlconjug3.readthedocs.io/en/latest/readme.html
[10] La documentación oficial de esta biblioteca está disponible en: https://pypi.org/project/pyspellchecker/
[11] Por ejemplo: https://www.ejemplos.co/perifrasis-verbal
[12] La colección de oraciones empleadas en las pruebas unitarias está disponible en: https://zenodo.org/records/17648792

CC BY NC ND

oraciones en español que contienen perífrasis verbales[13]. Los resultados obtenidos en este caso de uso fueron los siguientes:

- La prueba de concepto identificó y adaptó correctamente el 81,25% de las oraciones del conjunto.
- El 10,94% de los casos correspondieron a oraciones con perífrasis que no fueron detectadas por el sistema.
- En el 6,25% de los casos, la prueba de concepto detectó la perífrasis, pero no realizó la adaptación.
- En el 1,56% de los casos, la adaptación generada no tuvo sentido gramaticalmente.

La proporción de aciertos obtenida (81,25%) presenta un intervalo de confianza del 95% entre el 70,3% y el 89,7%, que indica que el rendimiento del sistema se mantiene claramente por encima del azar. De hecho, la prueba exacta binomial (H_0: $p = 0,5$) muestra que esta proporción es estadísticamente significativa ($p \approx 3,2 \cdot 10^{-6}$), lo que confirma que el método produce adaptaciones correctas de forma sistemática incluso con un conjunto reducido de oraciones.

Cabe señalar que, en los casos en los que el sistema no realizó una adaptación a lectura fácil, el 75% de las oraciones correspondían a perífrasis de tipo terminativo. Este patrón – de interés lingüístico – revela que la complejidad no deriva únicamente de la identificación sintáctica, sino también de la interpretación aspectual que estas construcciones codifican. Las perífrasis terminativas combinan información gramatical y pragmática que depende fuertemente del contexto discursivo, lo que explica por qué las reglas declarativas resultan insuficientes. Este hallazgo no es anecdótico: constituye un punto de partida para justificar la pertinencia del enfoque híbrido presentado en la sección 3.4.

3.3. Validación de la prueba de concepto por parte de usuarios finales

Para validar la prueba de concepto descrita en la sección anterior, se llevó a cabo una primera validación basada en un estudio con usuarios mediante cuestionarios en línea. Las personas participantes en esta validación inicial fueron personas con discapacidad cognitiva. El cuestionario[14] constaba de dos partes principales: (1) una sección dedicada a recoger información demográfica, trayectoria y experiencia de los participantes; y (2) otra parte con 13 preguntas cerradas de opción única, destinadas a conocer la opinión de los participantes sobre el grado de comprensión de las adaptaciones generadas por nuestra prueba de concepto en lo relativo a las perífrasis verbales.

La estructura de las 13 preguntas es homogénea. Cada una presenta dos frases: una original y otra adaptada automáticamente. Se pedía a los participantes que leyeran con calma cada par de frases y seleccionaran una de las siguientes opciones: (a) Entiendo mejor la primera frase; (b) Entiendo mejor la segunda frase; (c) Entiendo

[13] El conjunto de oraciones empleado en este caso de uso está disponible en: https://zenodo.org/records/17648792

[14] Un ejemplo del cuestionario en formato PDF está disponible en: https://zenodo.org/records/17648792

CC BY NC ND

bien ambas frases; o (d) No entiendo ninguna de las dos frases. El cuestionario fue evaluado por una experta en lectura fácil.

Un total de 22 personas participaron en la cumplimentación del cuestionario. De ellas, 9 eran hombres, 11 mujeres, 1 persona no binaria y 1 persona prefirió no declarar su género. Todas residían en Madrid. En cuanto al nivel de comprensión lectora, 7 personas manifestaron tener un nivel alto, 11 un nivel medio y 3 un nivel bajo (una persona no conocía su nivel). Por edades, 4 personas tenían entre 18 y 30 años, 13 entre 31 y 45, y 5 entre 46 y 60. En relación con la tipología de discapacidad, 18 personas presentaban discapacidad intelectual, 2 discapacidad intelectual y física, y otras 2 combinaban discapacidad intelectual y problemas de salud mental. En cuanto a su ocupación, 7 eran usuarios de centros ocupacionales, 12 trabajaban como validadores de lectura fácil, y el resto tenía ocupaciones diversas, incluyendo una persona desempleada, una jubilada y una responsable de kits.

Somos conscientes de que la principal limitación de esta actividad de validación es el tamaño reducido de la muestra. Sin embargo, siguiendo a Caine (2016), una muestra de 22 personas puede considerarse habitual en estudios de estas características. Además, conviene recordar que el proceso de reclutamiento es especialmente complejo cuando se trabaja con poblaciones minoritarias (Bills y Mills, 2022; Lennox et al., 2005), como es el caso de las personas con discapacidad cognitiva.

En la Tabla 2 se muestran los porcentajes de respuesta para cada opción en cada una de las preguntas (P) formuladas en el cuestionario de validación. Exceptuando la pregunta 2, en la mayoría de los casos la opción más elegida por los participantes fue la frase adaptada por el sistema. Esto también se observa en las preguntas más complejas, es decir, aquellas en las que existe un porcentaje no nulo en la opción "Ninguna". Como ejemplo, en la pregunta 7, 20 participantes seleccionaron la frase adaptada como la más comprensible, y solo 2 personas indicaron no haber comprendido ninguna de las dos.

Si consideramos como éxito el hecho de que el participante elija la frase adaptada como la más comprensible, y como fracaso cualquiera de las otras tres respuestas, se pueden realizar pruebas exactas binomiales para determinar si la proporción de éxitos en cada pregunta es significativamente superior al 50 %. En aquellas preguntas en que no se puede rechazar la hipótesis nula, no existe evidencia suficiente para afirmar que la adaptación mejora la comprensión respecto a la frase original. Sin embargo, en los casos donde sí se rechaza, puede concluirse que la adaptación facilita la comprensión. En concreto, las preguntas con más de 15 respuestas en la opción "Adaptada" (lo que equivale a una proporción de acierto mayor al 68,2 %) presentan un valor p inferior al nivel nominal de significación del 0,05. Así, se obtiene evidencia sólida sobre la eficacia de la herramienta en dichas preguntas. En cambio, en las preguntas P1, P2, P4, P8 y P13 no se rechaza la hipótesis nula, por lo que no puede afirmarse que la adaptación supere a la redacción original.

Tabla 2. Porcentajes de respuestas en la validación con usuarios

Opción	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
Original	9,1	81,8	9,1	9,1	4,5	0	0	9,1	4,5	9,1	18	4,5	31,8
Adaptada	54,5	4,5	90,9	68,2	86,4	95,5	90,9	63,6	86,4	81,8	64	77,3	54,5
Ambas	31,8	9,1	0	22,7	9,1	4,5	0	22,7	0	4,5	18	18,2	4,5
Ninguna	4,5	4,5	0	0	0	0	9,1	4,5	9,1	4,5	0	0	9,1

CC BY NC ND

3.4. Enfoque inicial basado en grandes modelos de lenguaje (LLM)

En la evaluación de nuestro prototipo, observamos que ciertas adaptaciones deberían depender del contexto, como ocurre con las perífrasis tempo-aspectuales terminativas. Este tipo de perífrasis no se adapta a una versión de lectura fácil mediante nuestro enfoque basado en reglas, como se menciona en la sección correspondiente. Esto indica que dicho método requiere un complemento para resultar válido en todos los casos.

Para afrontar este reto, recurrimos a los modelos transformadores (Vaswani et al., 2023), conocidos por su capacidad excepcional para captar matices del lenguaje natural a través del contexto. Gracias a los mecanismos de atención, estos modelos ponderan la relevancia contextual, permiten un aprendizaje flexible y procesan eficientemente los datos. Su arquitectura resulta especialmente adecuada para gestionar sutilezas lingüísticas, adaptarse a expresiones en evolución y mejorar progresivamente con más datos. Por tanto, decidimos aprovechar sus capacidades para lograr adaptaciones lingüísticas más precisas y sensibles al contexto.

Para ello, obtuvimos 172 oraciones en español con perífrasis terminativas mediante ChatGPT[15]. A fin de conformar un corpus paralelo[16] de oraciones originales y sus versiones adaptadas esperadas, realizamos manualmente la adaptación de dichas frases. De estas, 154 se emplearon para el entrenamiento y 18 para la validación. Intentamos usar LLM entrenados con datos en español como Falcon[17] o Mistral AI[18], pero dadas las necesidades computacionales que implicaban, comenzamos usando el modelo bart-large de Facebook (Lewis et al., 2019). Sin embargo, su entrenamiento en inglés redujo el rendimiento en tareas en español, generando incluso palabras inexistentes.

Tras obtener más recursos computacionales, cambiamos a modelos específicos del español como llama-2-7b-alpaca-es[19]. Este modelo fue elegido por ser viable en nuestros sistemas y por haber sido entrenado con corpus en español. Durante el proceso de ajuste fino del modelo, empleamos la técnica LoRA (*Low-Rank Adaptation*; Hu et al., 2021) que permite adaptar grandes modelos de lenguaje mediante la inserción de matrices de bajo rango en determinadas capas del modelo, sin necesidad de modificar todos sus parámetros. Esta técnica resulta especialmente útil en contextos con recursos computacionales limitados, ya que reduce significativamente el coste de entrenamiento. En nuestro caso, aplicamos LoRA durante 10 épocas, lo que implica que el modelo recorrió diez veces el conjunto de entrenamiento con el fin de optimizar sus parámetros adaptativos. La tasa de aprendizaje es otro concepto fundamental en el entrenamiento de modelos de aprendizaje automático. Esta tasa determina el tamaño de los pasos que el modelo da durante su proceso de aprendizaje. En nuestro experimento se eligió una tasa de

[15] https://chatgpt.com/
[16] El corpus paralelo está disponible en: https://zenodo.org/records/17648792
[17] La página oficial del modelo está disponible en: https://falconllm.tii.ae/falcon-40b.html
[18] La página oficial del modelo está disponible en: https://mistral.ai/products/la-plateforme#models
[19] La información y documentación del modelo está disponible en: https://huggingface.co/4i-ai/Llama-2-7b-alpaca-es

CC BY NC ND

aprendizaje de 0.0005[20], lo que implica que, en cada actualización, los parámetros del modelo se ajustaron en una magnitud relativamente pequeña. Este ajuste cuidadoso permite encontrar el conjunto óptimo de parámetros que minimiza los errores sin excederse en las correcciones.

Evaluamos el rendimiento del experimento inicialmente con la métrica de pérdida, que cuantifica la diferencia entre las predicciones del modelo y los datos reales. Obtuvimos una evaluación de pérdida de 0,6655, que refleja el comportamiento del modelo con los datos de validación, y una pérdida de 0,221 con los datos de entrenamiento. Aun así, estas métricas no bastaron para evaluar adecuadamente la calidad de las adaptaciones.

Dado que el objetivo del modelo era adaptar perífrasis tempo-aspectuales terminativas preservando su significado, decidimos usar la métrica BLEU (Papineni et al., 2002), originalmente concebida para evaluar traducciones automáticas. En nuestro conjunto de validación, obtuvimos un valor BLEU medio de aproximadamente 87,58.

De las 18 frases de validación, se adaptaron 17, lo que evidencia la eficacia de los modelos transformadores para afrontar desafíos complejos de adaptación lingüística en español.

4. CONCLUSIONES Y TRABAJO FUTURO

Las perífrasis verbales no suelen abordarse como un fenómeno específico que afecte a la comprensión lectora en español. Sin embargo, como se ha señalado a lo largo de este trabajo, constituyen un paradigma complejo desde el punto de vista sintáctico. Por tanto, resulta pertinente proponer un enfoque automático que contribuya a reducir esta complejidad sintáctica y, con ello, a facilitar la comprensión de los textos escritos.

La metodología de lectura fácil (LF) recomienda evitar este tipo de estructuras lingüísticas debido a su complejidad. En esta línea, el presente trabajo ha propuesto un método declarativo para la identificación y adaptación automática de perífrasis verbales en español, conforme a los principios de la LF. En relación con este objetivo, los resultados obtenidos permiten extraer varias conclusiones. En primer lugar, desde una perspectiva lingüística, mostramos que determinadas perífrasis – especialmente las tempo-aspectuales y las temporales – introducen cargas interpretativas que afectan de manera significativa a la comprensión lectora. Los patrones declarativos diseñados permiten capturar parte de esta complejidad y generar las adaptaciones de la estructura verbal, lo que abre la puerta a formalizar de manera sistemática pautas que hasta ahora se aplicaban únicamente de forma manual. En segundo lugar, los resultados de la prueba de concepto sugieren que el enfoque basado en reglas es capaz de identificar y adaptar una proporción considerable de casos, aunque presenta limitaciones importantes. El hallazgo de que el 75% de los errores se concentre en las perífrasis terminativas aporta una evidencia clara de los límites del método simbólico, al tratarse de construcciones cuya

[20] Las cifras presentadas en esta sección siguen las convenciones ortotipográficas del inglés, dado que esta es la lengua habitual en el desarrollo de este tipo de implementaciones.

CC BY NC ND

interpretación depende del contexto discursivo. Esto permite establecer un vínculo entre los resultados y la teoría lingüística sobre el aspecto verbal.

En tercer lugar, los experimentos preliminares con modelos neuronales muestran que los LLM pueden complementar estas limitaciones, especialmente en estructuras donde la interpretación no se deriva de la superficie sintáctica. El rendimiento observado, aun con un corpus pequeño, confirma la pertinencia de un enfoque híbrido que combine reglas lingüísticas con métodos subsimbólicos. Finalmente, la validación con personas con discapacidad cognitiva apunta a que las adaptaciones generadas facilitan la comprensión en un número significativo de casos, aunque el tamaño reducido y la composición de la muestra exigen interpretar estos resultados con cautela. En conjunto, el estudio aporta una primera aproximación sistemática al tratamiento automático de las perífrasis verbales en textos accesibles. Futuras investigaciones deberán ampliar los corpus, profundizar en la integración de modelos híbridos y realizar evaluaciones con muestras más amplias y diversas para consolidar las conclusiones obtenidas. En esta dirección, resulta necesario reforzar la validación inicial incorporando la participación de personas expertas en lectura fácil, con el fin de llevar a cabo una validación cruzada de las adaptaciones generadas por el método. Del mismo modo, conviene seguir explorando la combinación entre reglas lingüísticas y modelos de lenguaje a gran escala, aplicando este enfoque híbrido a un repertorio más general de perífrasis verbales.

REFERENCIAS

AENOR. 2018. *Lectura Fácil. Pautas y recomendaciones para la elaboración de documentos* (UNE 153101:2018 EX). Asociación Española de Normalización.

Bills, K. L. y B. Mills. 2022. Limitations when conducting quantitative disability research. *Journal of Research Initiatives* 6(2): Artículo 10.

Bott, S., L. Rello, B. Drndarevic y H. Saggion. 2012. Can Spanish be simpler? LexSiS: Lexical Simplification for Spanish. En *Proceedings of COLING 2012*, Mumbai, 357–374.

Caine, K. E. 2016. Local standards for sample size at CHI. En *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.

Castro-Sánchez, N. A., I. C. Domínguez, G. Sidorov y A. M. Rebollar. 2015. Hacia una extracción automática de colocaciones en definiciones de verbos de un diccionario explicativo en español. *Revista Signos: Estudios de Lingüística* 88: 174–196.

Celayeta Gil, N. 2011. El tratamiento didáctico de las perífrasis verbales en los materiales dedicados a la enseñanza del E/LE y 2L en la Red. En C. Hernández González, A. Carrasco Santana y E. Álvarez Ramos, coord. *La red y sus aplicaciones en la enseñanza-aprendizaje del español como lengua extranjera*. Asociación para la Enseñanza del Español como Lengua Extranjera – ASELE, 173–184.

Díez, E., N. Rodríguez, A. Fernández, M. A. Alonso, M. A. Díez-Álamo, M. J. Sánchez y D. Wojcik. 2019. *E2R-Helper [Asistente Lectura Fácil]*. Instituto Universitario de Integración en la Comunidad – Universidad de Salamanca.

Drndarevic, B., S. Stajner y H. Saggion. 2012. Reporting simply: A lexical simplification strategy for enhancing text accessibility. En *Proceedings of Easy-to-Read on the Web Symposium*. W3C.

García González, J. 1992. *Perífrasis verbales*. Sociedad General Española de Librería.

García Muñoz, Ó. 2012. *Lectura fácil: Métodos de redacción y evaluación*. Real Patronato sobre Discapacidad.

CC BY NC ND

Gil, N. C. 2011. El tratamiento didáctico de las perífrasis verbales en los materiales dedicados a la enseñanza del E/LE y 2L en la red. En *La Red y sus aplicaciones en la enseñanza-aprendizaje del español como lengua extranjera*, 173–184.
Hearst, M. A. 1992. Automatic acquisition of hyponyms from large text corpora. En *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*, 539–545.
Holz, F. y C. Biemann. 2008. Unsupervised and knowledge-free learning of compound splits and periphrases. En *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'08)*. Springer, 117–127.
Hu, E. J., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang y W. Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint*, arXiv:2106.09685.
Iglesias, A., I. Cobián, A. Campillo, J. Morato y S. Sánchez-Cuadrado. 2020. Comp4text checker: An automatic and visual evaluation tool to check the readability of Spanish web pages. En *Computers helping people with special needs*. Springer, 258–265.
Inclusion Europe. 2009. *Information for all. European standards for making information easy to read and understand*. Inclusion Europe.
Lennox, N., M. Taylor, T. Rey-Conde, C. Bain, D. M. Purdie y F. Boyle. 2005. Beating the barriers: Recruitment of people with intellectual disability to participate in research. *Journal of Intellectual Disability Research* 49(4): 296–305.
Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov y L. Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.
Mata, F. S., J. L. G. Ortega y C. G. Mieres. 2007. Habilidades lingüísticas y comprensión lectora. Una investigación empírica. *Bordón: Revista de Pedagogía* 59: 153–166.
Mera, J. J. A. 2015. *Representación computacional de las perífrasis de fase: De la cognición a la computación* [tesis doctoral]. Universitat de Barcelona.
Moreno, L., R. Alarcón y R. Martínez. 2020. EASIER System. Language resources for cognitive accessibility. En *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '20)*. Association for Computing Machinery.
Nomura, M., G. S. Nielsen, IFLA y Library Services to People with Special Needs Section. 2010. *Guidelines for easy-to-read materials*. IFLA Headquarters.
Papineni, K., S. Roukos, T. Ward y W.-J. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. En *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*, 311–318.
RAE, ASALE. 2009. *Nueva gramática de la lengua española*. Espasa Calpe.
Rello, L., R. Baeza-Yates y H. Saggion. 2013. DysWebxia: Textos más accesibles para personas con dislexia. *Procesamiento del Lenguaje Natural* 51: 205–208.
Russell, S. y P. Norvig. 2021. *Artificial Intelligence, global edition: A modern approach*. Pearson Deutschland.
Saggion, H. 2022. Text simplification. En R. Mitkov ed. *The Oxford handbook of computational linguistics*. Oxford University Press, 1114–1136.
Saggion, H., S. Stajner, S. Bott, S. Mille, L. Rello y B. Drndarevic. 2015. Making it Simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Transactions on Accessible Computing* 6(4): Artículo 14.
Suárez-Figueroa, M. C., E. Ruckhaus, J. López-Guerrero, I. Cano y Á. Cervera. 2020. Towards the assessment of easy-to-read huidelines using Artificial Intelligence techniques. En *Computers Helping People with Special Needs (ICCHP 2020)*, Lecture Notes in Computer Science 12376, Springer, 74–82.
Suárez-Figueroa, M. C., I. Diab, E. Ruckhaus e I. Cano. 2022. First steps in the development of a support application for easy-to-read adaptation. *Universal Access in the Information Society*. https://doi.org/10.1007/s10209-022-00946-z
Torrego, L. G. 1988. *Perífrasis verbales. Sintaxis, semántica y estilística*. Arco Libros.

CC BY NC ND

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser e I. Polosukhin. 2023. Attention is all you need. *arXiv preprint*, arXiv:1706.03762.

CC BY NC ND