RLA. Revista de Lingüística Teórica y Aplicada
Concepción (Chile), 62 (2), II Sem. 2024, pp. 37-62.

ISSN 0033-698X (impresa)
ISSN 0718-4883 (on-line)
https://doi.org/10.29393/RLA62-2CTYR20002

COMPARACIÓN ENTRE TEXTOS NATURALES Y TEXTOS SINTÉTICOS DEL GÉNERO TESIS EN FUNCIÓN DE VARIABLES DISCURSIVAS[1]

DISCOURSE-BASED COMPARISON OF NATURAL AND SYNTHETIC TEXTS OF THE THESIS GENRE

YVONE LAINES RUIZ
Pontificia Universidad Católica de Valparaíso
yvone.laines@pucv.cl
ORCID: https://orcid.org/0009-0006-5984-7291

ROGELIO NAZAR
Pontificia Universidad Católica de Valparaíso
rogelio.nazar@pucv.cl
ORCID: https://orcid.org/0000-0002-8853-1353

RESUMEN

La irrupción de la inteligencia artificial generativa presenta oportunidades y desafíos para las prácticas establecidas de escritura académica. En este contexto, el presente artículo propone un estudio descriptivo de las diferencias existentes, en cuanto a variables discursivas, entre textos naturales y artificiales en un corpus de tesis de licenciatura y doctorado en las disciplinas de acuicultura, derecho y lingüística. Se utilizó una metodología de lingüística de corpus para medir variables como longitud de párrafos y oraciones, riqueza léxica, marcadores discursivos, deixis y modalizadores. Los principales hallazgos indican que los textos naturales presentan patrones de longitud característicos, así como mayor riqueza léxica, frecuencia y diversidad en el uso de recursos discursivos en comparación con los textos sintéticos. En particular, los textos naturales muestran mayor cantidad de marcas de subjetividad, mientras que los textos sintéticos muestran un uso frecuente de determinados marcadores estructuradores, menor diversidad de mecanismos y una estructura más uniforme.

Palabras clave: alfabetización académica, inteligencia artificial generativa, lingüística de corpus, tesis, variables discursivas.

[1]Esta investigación ha sido financiada por el Proyecto Fondecyt Regular 1231594, titulado "Mapa de las metáforas conceptuales en sustantivos y verbos del español: un estudio de los patrones metafóricos basado en corpus" (años 2023-2027), dirigido por Irene Renau y con el segundo autor de este artículo como coinvestigador.

ABSTRACT

The irruption of generative artificial intelligence poses challenges and opportunities for established academic writing practices. In this context, this paper proposes a descriptive study of the differences, in discourse variables, between natural and synthetic texts in undergraduate and doctoral theses in disciplines such as aquaculture, law and linguistics. A corpus linguistics methodology was used to measure paragraph and sentence length, lexical richness, discourse markers, deixis and modality. The main findings indicate that natural texts exhibit characteristic length patterns, greater lexical richness, as well as more frequency and diversity in the use of discourse resources compared to synthetic texts. In particular, natural texts present greater concentration of subjectivity markers, while the synthetic ones show a frequent use of certain structuring markers, less variability of mechanisms, and a more uniform structure.

Keywords: academic literacy, generative artificial intelligence, corpus linguistics, theses, discourse variables.

Recibido: 03/08/2024 Aceptado:10/11/2024

1. INTRODUCCIÓN

La alfabetización académica constituye uno de los múltiples desafíos que enfrentan los estudiantes al ingresar a la educación superior, ya que este proceso implica la inserción gradual en una comunidad disciplinar a la que aspiran pertenecer (Carlino, 2013; Navarro, 2017; Marinkovich et al., 2018). Para lograr esta inserción, los estudiantes deben desarrollar competencias específicas de lectura y escritura propias de cada disciplina (Velásquez & Marinkovich, 2016), las cuales son fundamentales para su éxito académico. Una de las formas en que los estudiantes pueden demostrar el logro de estas competencias durante su formación es a través de la presentación de la tesis, tanto en pregrado como en posgrado (Parodi et al., 2008; González & Ibáñez, 2017). Este género académico complejo requiere que los estudiantes dominen las convenciones discursivas y epistemológicas de su disciplina, convirtiéndose así en una manifestación concreta de su alfabetización académica, la cual varía según su grado (Meza & Rivera, 2018). En este contexto, la irrupción de las tecnologías de Inteligencia Artificial Generativa (IAG) presenta oportunidades y desafíos para el normal desenvolvimiento de este proceso en la carrera universitaria (Crompton & Burke, 2023). En este artículo utilizamos el término *texto sintético* para denominar el texto generado automáticamente por estas tecnologías y que, por sus características, parece tan similar al texto natural que resulta difícil advertir que no ha sido producido por humanos.

Ante este escenario, no es sorprendente que el uso de una tecnología tan poderosa como la IAG genere debate en varios campos. En la educación superior,

Popenici y Kerr (2017) señalan que la tecnología e inteligencia artificial están cambiando la forma de enseñar, aprender y organizar las universidades. Algunos expertos creen que la IAG puede ser útil para mejorar la enseñanza universitaria (Zawacki-Richter et al., 2019; Hwang et al., 2020; Holmes & Tuomi, 2022; García-Peñalvo et al., 2024). Otros, en cambio, plantean que si bien puede ser un aporte para la redacción y supervisión de tesis, implica algunos riesgos, entre los que se encuentra la posibilidad de fraude, cuando se entregan como propios trabajos desarrollados por IAG (Zawacki-Richter et al., 2019; Borger et al., 2023). El debate ha motivado recientes investigaciones en el desarrollo de herramientas para identificar texto académico redactado con IAG, como ZeroGPT[2] o el identificador Open AI Classifier[3] (actualmente no disponible), entre otros (Desaire et al., 2023).

El presente artículo expone los primeros resultados de un trabajo en curso sobre las características discursivas del texto sintético. Con el objetivo de identificar posibles diferencias en el plano discursivo entre textos naturales y sintéticos en el discurso académico y, en particular, en el género tesis, el presente estudio considera la siguiente pregunta de investigación: ¿Qué diferencias existen, en cuanto a variables discursivas, entre textos naturales y textos sintéticos en la escritura académica? Para ello, se trabajó con la metodología de la lingüística de corpus para medir algunas variables discursivas elementales tales como medidas de longitud de oración y párrafo, medidas de diversidad léxica, marcadores discursivos y marcadores de subjetividad (deixis y modalización) en un corpus conformado por 45 tesis de licenciatura y doctorado.

El trabajo ofrece algunos aportes relevantes para el tema desde el enfoque de la lingüística aplicada. En primer lugar, establece parámetros lingüísticos objetivos para comparar textos naturales y textos sintéticos, lo cual es fundamental para salvaguardar la integridad académica en un contexto donde la IAG ha transformado, como indicábamos, las prácticas de escritura en la universidad. En segundo lugar, las variables discursivas seleccionadas permiten comprender las mecánicas subyacentes de los modelos de lenguaje generativo y de qué manera reproduce o se diferencia de las prácticas de escritura humana. Finalmente, esta investigación representa también un aporte para la alfabetización académica, al identificar algunas características distintivas del discurso natural. Si bien se trata de un estudio puramente descriptivo, aporta datos que pueden ser útiles para futuras investigaciones que busquen clasificar textos en las dos categorías de natural y sintético.

[2] https://www.zerogpt.com/

[3] https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/

2. MARCO TEÓRICO

2.1 Alfabetización académica y el género discursivo tesis en la educación superior

En las últimas décadas se han llevado a cabo investigaciones en torno a las problemáticas de los estudiantes que ingresan a la educación terciaria en las prácticas discursivas. Carlino (2003) planteó el tema de la alfabetización académica como objeto de estudio relevante en el ámbito de la educación superior, definida como la responsabilidad individual de cada alumno de acceder al conjunto de nociones y estrategias requeridas para la inserción en la cultura discursiva de una disciplina particular. Estos estudios han motivado esfuerzos de las instituciones de educación superior por enseñar las prácticas para participar de los géneros de cada campo de estudio (Carlino, 2013).

Navarro (2017) destaca el interés del desarrollo de las literacidades académicas por su función epistémica, retórica, habilitante, empoderadora y expresiva, ya que permite que el alumno aprenda, se comunique y exprese los conocimientos adquiridos. Además, según Velásquez y Marinkovich (2016), la alfabetización académica difiere entre disciplinas. Su estudio, que comparó las licenciaturas en Historia y Biología, reveló que aunque comparten tres momentos esenciales en la consecución de la escritura académica, cada disciplina requiere un tipo específico de pensamiento (histórico o científico) y utiliza géneros discursivos distintos. El acceso a este conocimiento disciplinar considera la comprensión y producción de diversos géneros académicos, siendo la tesis el género discursivo que permite dar cuenta de su inserción en la comunidad discursiva en el ámbito científico (Parodi et al., 2008; González & Ibáñez, 2017).

Para los estudiantes, naturalmente, la tesis es un desafío porque es la instancia en que deben acreditar sus conocimientos y habilidades para la obtención de un grado académico como licenciado o magíster. En el caso del doctorado, las exigencias son todavía más rigurosas, ya que se espera de ellos no solo que demuestren conocimiento de la disciplina, sino también que sean capaces de aportar conocimiento original y relevante (Venegas et al., 2016; Meza & Rivera, 2018; Rey & Velásquez, 2023).

2.2 IAG en la escritura académica de educación superior

El origen de la Inteligencia Artificial (IA) puede situarse aproximadamente en la década de 1950, cuando John McCarthy acuñó el término durante un taller en el Dartmouth College en 1956. La propuesta original de McCarthy planteaba la posibilidad de simular aspectos de la inteligencia humana mediante máquinas,

incluyendo el uso del lenguaje, la formación de conceptos y la resolución de problemas (Russell & Norvig 2010; Popenici & Kerr, 2017; Zawacki-Richter et al., 2019). Desde entonces, la definición de IA ha evolucionado, abarcando un amplio espectro de tecnologías y métodos. Baker y Smith (2019) la describen como computadoras realizando tareas cognitivas asociadas con la mente humana, especialmente el aprendizaje y la resolución de problemas. Este concepto engloba diversas áreas y técnicas como el aprendizaje automático, el procesamiento del lenguaje natural y las redes neuronales.

Como ya se mencionó, los últimos desarrollos de esta tecnología han generado creciente interés por su aplicación en el contexto de la educación. Crompton y Burke (2023) realizaron una revisión sistemática de las publicaciones sobre inteligencia artificial en educación superior (IAES) y destacaron que las investigaciones en el área han tenido un alza sostenida desde 2021. La mayor proporción de publicaciones en este ámbito proviene de China, principalmente dedicada al estudiante de pregrado (Crompton & Burke, 2023, p. 8). Este enfoque en el uso de la IA para apoyar a los estudiantes de pregrado se alinea con los hallazgos de Borger et al. (2023), quienes detallan múltiples herramientas y tecnologías de IA que ayudan tanto a estudiantes como a académicos en diversos aspectos de su trabajo, tales como la producción de investigaciones, la gestión de referencias, la lectura de artículos, el análisis de datos, la traducción, la colaboración y el diseño de encuestas, entre otros.

Entre las IAG, la que primero captó la atención mundial en materia de chatbots fue ChatGPT 3.5 (OpenAI, 2022), lanzada en noviembre de 2022. Este sistema se basa en un modelo lingüístico producido gracias a la arquitectura de redes neuronales conocida como transformador generativo preentrenado (*generative pretrained transformer*, GPT) (OpenAI, 2022), diseñado originalmente para la traducción automática (Vaswani et al., 2017) pero que, de manera sorprendente, resultó efectiva también para una gran diversidad de tareas para las cuales no había sido diseñado (Kojima et al., 2022). En su aplicación como chatbot, y gracias a la elaboración de grandes modelos de lengua (*large language models*, LLM), es capaz de proporcionar respuestas originales a partir de instrucciones en texto libre, con una fluidez y corrección gramatical jamás alcanzada en la historia de la lingüística computacional (Goldstein et al., 2022). El modelo de ChatGPT 3 se sustentó en 570 GB de datos, que comprenden 300.000 millones de palabras, de las cuales se derivan 175.000 millones de parámetros, es decir, valores aprendidos por el modelo durante el entrenamiento que determinan su comportamiento y capacidad para realizar tareas específicas (Brown et al., 2020; Bender et al., 2021).

El lanzamiento de este tipo de aplicaciones ha permeado gran parte de los ámbitos de la sociedad, incluyendo la educación, y se percibe a la vez como oportunidad y como amenaza. Entre las oportunidades, Sabzalieva y Valentini

(2023) indican que tienen un posible uso en la enseñanza y aprendizaje, la investigación, la administración y el compromiso con la comunidad. En el ámbito de la enseñanza-aprendizaje, destaca su rol como un generador de respuestas alternativas, oponente socrático[4], asesor de trabajo colaborativo, guía complementaria y evaluador dinámico (Katinskaia & Yangarber, 2024). La variedad de posibles tareas parece inagotable. En la investigación puede aportar en las diversas fases del proceso como en el diseño, recogida y análisis de datos, redacción de textos y traducción automática. Asimismo, podría apoyar en tareas administrativas de la educación superior o resolver dudas de los estudiantes.

Finalmente, en el área de participación en la comunidad, la IAG posibilita la creación de propuestas adecuadas a diferentes contextos específicos. Borger et al. (2023) indican usos potenciales de esta tecnología, entre los que destacan la escritura y lectura de becas y trabajos; la superación de barreras lingüísticas para hablantes no nativos de inglés en ciencias; la organización de conferencias; análisis, codificación e interpretación de datos; análisis de imágenes complejas y mejoras en la experiencia de redacción de tesis para estudiantes y supervisores.

Los riesgos, por su parte, son diversos, e incluyen discriminación de género y diversidad, sesgos, falta de regulación, falta de protección de datos y amenaza a la integridad académica (Bender et al., 2021). El último punto, especialmente, representa una de las grandes preocupaciones de la academia y ha motivado el desarrollo de herramientas para detectar la creación de texto sintético, como ZeroGPT y Open AI Classifier, aunque los resultados todavía no alcanzan suficiente precisión (Desaire et al., 2023).

El desarrollo de técnicas para la detección de textos sintéticos se perfila como una vía de investigación emergente. Desaire et al. (2023) describen un método para detectar artículos sintéticos en el área de química, para el que informan resultados prometedores. Utilizaron 20 medidas para la detección de textos sintéticos, entre las que destacan, como valor predictor, la variación de la longitud de las oraciones. Este hallazgo motiva a continuar explorando esa y otras variables discursivas como la longitud de párrafo, el uso de marcadores discursivos y las marcas de subjetividad.

2.3 Variables discursivas básicas

En este breve estudio se explora el uso de algunas de las variables discursivas que podrían servir para reconocer textos generados por IA. En este sentido, se entiende por variables discursivas aquellos elementos que influyen en la composición y comprensión de un discurso (Calsamiglia & Tusón, 1999). Para ello, se revisará

[4] En este caso, el sistema permite al estudiante utilizar el método socrático a través de una discusión con la inteligencia artificial, con la finalidad de desarrollar el pensamiento crítico.

brevemente la conceptualización de las variables seleccionadas, que corresponde a medidas de longitud de enunciado, de diversidad léxica, marcadores discursivos, deixis y modalizadores.

Medidas de longitud de enunciado

Las medidas de longitud de enunciado (Nazar, 2024) se aplican al análisis de la extensión de oraciones, párrafos o bien textos contando sus palabras, letras o morfemas, y han sido utilizadas en diversos contextos. Una de las primeras aplicaciones que se encontró fue como medida de complejidad sintáctica (Flesh, 1949; Brown, 1973), que es aplicable, a su vez, al estudio del desarrollo lingüístico en niños, entre otras posibilidades. La industria editorial, por su parte, también se ha interesado por la aplicación de este tipo de medidas para el desarrollo y clasificación de materiales de aprendizaje (Graesser et al., 2004). Además, otros investigadores han explorado la relación que existe entre esta variable y las tipologías textuales (Kelih et al., 2006). En la actualidad, como ya mencionamos, se está produciendo una renovación del interés por este tipo de medidas para el análisis lingüístico de textos sintéticos (Desaire et al., 2023), ya que podría ser una clave para identificarlos.

Medidas de diversidad léxica

Tradicionalmente, en los estudios lexicológicos se ha empleado como medida de riqueza o diversidad léxica de un texto lo que se conoce como el *type-token ratio*, definido como la proporción entre vocabulario, o cantidad de palabras distintas, y extensión, es decir, la cantidad de palabras totales (Baayen, 2008). Esta medida ha sido criticada por excesivamente simplista, ya que no es independiente del tamaño de los textos (Herdan, 1964). Sin embargo, continúa siendo utilizada en la actualidad (Rojo, 2021), justamente debido a su facilidad de implementación, y por el hecho de continuar siendo válida siempre cuando la extensión de los textos que se comparan sea una variable controlada.

Marcadores discursivos

Los marcadores del discurso (MD) son unidades lingüísticas que sirven para enlazar proposiciones, estructurar la información del texto o bien para dirigir los intercambios comunicativos (Robledo, 2021; Robledo & Nazar, 2023). Los MD son invariables y no cumplen una función sintáctica en la oración. Su función es, en cambio, discursiva, ya que sirven como guía de las inferencias que el lector debe hacer para interpretar el texto (Martín Zorraquino & Portolés, 1999). De este modo, facilitan el establecimiento de las relaciones semánticas entre diferentes partes del texto, ya sean estas enunciados individuales o secuencias más extensas (Calsamiglia & Tusón, 1999).

Las palabras y expresiones que se utilizan actualmente como MD provienen de diversas categorías gramaticales, tales como conjunciones, locuciones conjuntivas, adverbios, locuciones adverbiales, interjecciones, expresiones performativas, entre otras (Calsamiglia & Tusón, 1999; Robledo, 2021; Robledo & Nazar, 2023). Es tal la heterogeneidad gramatical que presentan, que lo que determina su unidad como categoría es, entonces, únicamente la función que cumplen en el discurso.

El tema ha suscitado el interés de diversos lingüistas particularmente en las últimas décadas. En el caso de la lengua castellana destacan, entre otras, las propuestas de Casado Velarde (1993), Calsamiglia y Tusón (1999), Martín Zorraquino y Portolés (1999), Montolío (2001) y Fuentes Rodríguez (2003). Casado Velarde (1993) propone las categorías de adverbios modificadores oracionales y los marcadores de función textual. Calsamiglia y Tusón (1999) proponen las categorías de marcadores de ordenación, marcadores de operaciones discursivas, conectores y marcadores interactivos. Martín Zorraquino y Portolés (1999) establecen cinco grandes grupos de MD según sus funciones: estructuradores de la información, conectores, reformuladores, operadores argumentativos y marcadores conversacionales. Montolío (2001) se concentra en el estudio de conectores, y distingue entre conectores opositivos/contraargumentativos, causales/consecutivos y aditivos. Fuentes Rodríguez (2003), por su parte, utiliza las categorías de conectores y operadores discursivos.

Existen también otros tipo de categorización más centrada en la función del marcador, entre los que destacan la propuesta de Cortés Rodríguez (2001), quien diferencia entre los marcadores de relación de los constituyentes textuales y los marcadores de estructuración de la conversación; Cortés Rodríguez y Camacho (2005), quienes proponen una clasificación entre marcadores interactivos y los marcadores textuales y, finalmente, Briz et al. (2008), quienes distinguen las funciones de conexión, modalización, focalización y control del contacto. Al respecto de cualquier clasificación basada en la función, conviene señalar una característica adicional de los MD que es su polifuncionalidad, es decir, la capacidad que puede tener un mismo MD para cumplir distintas funciones según el contexto (Martín Zorraquino & Portolés, 1999; Robledo, 2021).

Deixis

El concepto de deixis refiere a elementos lingüísticos cuyo significado depende directamente del contexto (Lozano et al., 1989; Calsamiglia & Tusón, 1999; Cuenca, 2010). La importancia de los deícticos radica en su capacidad para situar a los participantes de una comunicación en relación con los elementos que los rodean, permitiendo una comprensión completa del texto. El sentido de los deícticos se completa al entender quién emite el mensaje, a quién se dirige, y en qué tiempo y lugar ocurre la situación comunicativa. Estos datos externos al texto

(exofóricos) pueden condicionar la interpretación correcta de un texto ya que un mismo texto puede cambiar de contenido dependiendo de quién lo emite, a quién va dirigido, además de cuándo y dónde tuvo lugar la enunciación.

Existen tres categorías de deixis ampliamente aceptadas: personal, temporal y espacial. La deixis personal se refiere a elementos relacionados con los participantes del discurso, como pronombres de primera y segunda persona (*yo, tú/vos, nosotros, mi, tu,* etc.), así como los morfemas verbales de primera y segunda persona (*digo*, *conocí*, *recuerdas*, etc.).

La deixis temporal proporciona información sobre la simultaneidad, anterioridad o posterioridad respecto al momento de la enunciación (*mañana, ayer, pasado mañana*, etc.) y debe distinguirse del resto de marcas temporales que no refieren al momento de la enunciación sino a la temporalidad de un hecho relatado. Sería el caso de marcas tales como *entonces, la víspera, el día después*, etc., ya que estas refieren a algo dicho antes en el texto y establecen una relación endofórica en lugar de exofórica.

Finalmente, la deixis espacial indica la posición de elementos en relación con los participantes del evento comunicativo. Formas de deixis espacial pueden ser los demostrativos (*este, ese, aquel*, etc.), pronombres neutros (*esto, eso, aquello*, etc.), proformas adverbiales (*aquí, ahí, allí*, etc.), entre una diversidad de otras categorías, ya que también puede considerarse deíctica una importante proporción del vocabulario (Kerbrat-Orecchioni, 1997), como en el caso de verbos tales como *venir* o *regresar*, cuya utilización implica necesariamente un punto de referencia espacial por parte del hablante. Nuevamente, en el caso de los deícticos espaciales también cabe señalar que su función siempre depende del contexto, ya que la misma forma como, por ejemplo, *allí,* podría tener una función anafórica en lugar de deíctica, si está siendo utilizada como correferente, es decir, para recuperar un referente o tema mencionado antes.

Modalizadores

Los modalizadores, también denominados operadores modales, corresponden a marcas lingüísticas que expresan la subjetividad del hablante sobre su enunciado. Al igual que sucede con los marcadores discursivos, los operadores modales tampoco pueden delimitarse conceptualmente desde un punto de vista lingüístico a una sola categoría gramatical, ni a una clase cerrada de palabras. Los modalizadores abarcan una amplia gama de expresiones, desde elementos monolexemáticos hasta construcciones polilexemáticas (Calsamiglia & Tusón, 1999). Por lo tanto, se pueden definir como operadores de clase abierta que marcan un tipo de modalización en un enunciado (Obreque & Nazar, 2023).

Las tres categorías principales de los modalizadores provienen de la lógica clásica: epistémica, deóntica y alética. Los modalizadores epistémicos indican el

grado de compromiso y conocimiento del hablante (Pérez Canales, 2009). Verbos como *saber* y *creer, entre otros,* se han vinculado a esta categoría. Los modalizadores deónticos expresan obligación (Van Dijk, 2012), con expresiones como *se debe* o *es necesario que,* etc. Finalmente, los modalizadores aléticos son aquellos que manifiestan necesidad o posibilidad (Van Dijk, 2012), tales como *probablemente* o *es posible que,* etc. Sumado a las tres principales, se encuentran las categorías de modalizadores valorativos o axiológicos, que expresan valoraciones negativas o positivas respecto de un enunciado (Kerbrat-Orecchioni, 1997); modalizadores veredictorios, que señalan la veracidad de un enunciado (Greimas & Courtés, 1991); y los modalizadores volitivos, que expresan deseos o intenciones (García-Negroni & Tordesillas, 2001).

Diversos estudios han intentado identificar e inventariar los modalizadores (Pyatkin et al., 2021). En español, destacan proyectos lexicográficos que documentan, entre otras partículas, varios ejemplos de operadores modales, tales como el Diccionario de partículas de Santos Río (2003), el Diccionario de partículas discursivas del español (Briz et al., 2008) y el Diccionario de conectores y operadores del español (Fuentes Rodríguez, 2009). Además, se han producido intentos recientes por inventariar y clasificar los operadores modales en lengua castellana mediante su extracción automática a partir de corpus paralelos (Obreque & Nazar, 2023).

3. METODOLOGÍA

Con el propósito de responder a la pregunta de investigación, acerca de las posibles diferencias entre textos naturales y sintéticos en materia de variables discursivas, hemos procedido al análisis por medio de estadísticas descriptivas de una muestra de textos del género tesis. A continuación, describimos los materiales, los procedimientos y la medición de cada una de las variables en análisis.

3.1 Materiales

Se utilizó como muestra de datos el corpus compilado por Ignacio Lobos en su tesis doctoral (en preparación), gentilmente cedido para este estudio. El corpus (tabla I) consiste en una muestra no probabilística de tesis de pregrado (licenciatura) y de posgrado (doctorado), en las disciplinas de acuicultura, derecho y lingüística. Las tesis pertenecen a la Universidad de Chile (UCH), Pontificia Universidad Católica de Valparaíso (PUCV) y la Universidad Católica del Norte (UCN). Esta variedad beneficia la representatividad y el equilibrio de la muestra, al considerar diversas disciplinas y contextos del país y a la vez un número estable de tesis por disciplina y grado. Los documentos se encuentran en formato Word (.docx) y se identifican

con el siguiente código: abreviatura de *disciplina_inicial* de *grado_n°* tesis. Por ejemplo, *ACU_D_3* designa el tercer caso de tesis de acuicultura del grado de doctorado.

Tabla I. *Organización de las tesis proporcionadas.*

Grado	Disciplina	Universidad	Total
Doctorado	Acuicultura	1 UCH 1 PUCV 1 UCN 2 programa Cooperativo	5
Doctorado	Derecho	3 UCH (1 cotutela) 2 PUCV	5
Doctorado	Lingüística	5 PUCV	5
Licenciatura	Acuicultura	10 PUCV	10
Licenciatura	Derecho	2 UCH 8 PUCV	10
Licenciatura	Lingüística	3 UCH 7 PUCV	10
			45

Para la presente investigación, a partir de este corpus natural se procedió a crear un corpus sintético utilizando ChatGPT 3.5. El método para obtener este segundo corpus fue utilizar el título y el primer párrafo de la introducción de cada tesis como parte de la instrucción para generar artificialmente un segundo párrafo. Para esta instrucción se recurrió al estilo de la perspectiva profesional (Morales-Chan, 2023), en el que se solicita a la IAG adoptar un rol específico para realizar la tarea, es decir, considerando en la formulación el rol, el contexto, la tarea, el estilo y un ejemplo. Siguiendo este esquema, la instrucción fue la siguiente:

> *"Eres un estudiante universitario y estás escribiendo una tesis para optar al grado de (grado) en (disciplina). Te entregaré el título de la tesis y el primer párrafo de la sección de introducción para que escribas desde el segundo párrafo de esta introducción en adelante. Procura escribir con un registro académico. Título (...). Primer párrafo introducción (...)".*

Una vez obtenidos los datos, se organizaron en una matriz que identifica las tesis por su código, el grado, el título, el primer y segundo párrafo de la introducción natural y el párrafo sintético (tabla II). De esta matriz, a partir de las columnas "segundo párrafo natural" y "segundo párrafo sintético", se derivaron 6 archivos de texto necesarios para la investigación: 1) licenciatura natural, 2) licenciatura sintética, 3) doctorado natural, 4) doctorado sintético, 5) total natural y 6) total sintético.

Tabla II. Organización tabla de datos del corpus

TESIS	GRADO	TÍTULO	PRIMER PÁRRAFO ORIGINAL	SEGUNDO PÁRRAFO ORIGINAL	SEGUNDO PÁRRAFO SINTÉTICO
ACU_D_1	DOCTORADO	EFECTO DE LA COMPOSICIÓN DE ÁCIDOS GRASOS ALTAMENTE INSATURADOS n-3, DE LA DIETA DE [...].	El halibut del Atlántico (Hippoglossus hippoglossus) es una especie bentónica que habita fondos marinos [...].	Esta especie tiene un hábito alimenticio carnívoro. Su estómago se caracteriza por ser grande y su intestino corto, posee de 3 a 4 ciegos pilóricos[...].	La acuicultura del halibut del Atlántico ha experimentado un crecimiento significativo en las últimas décadas, convirtiéndose en una actividad económica [...].

3.2 Procedimientos

Ambas muestras de textos naturales y sintéticos fueron sometidos a la misma batería de medidas. La primera medición fue la longitud oracional y de párrafo, utilizando para ello un script en el lenguaje Perl[5] que aplica la función *length* para medir la extensión de texto en caracteres. Para poder llevar a cabo la medición de la longitud de oraciones fue necesario primero segmentar los párrafos en oraciones, tarea no trivial ya que los signos de puntuación normalmente utilizados para ello no tienen únicamente esa función. El caracter utilizado como punto final de oración puede aparecer también por ejemplo en abreviaturas y números, por lo que cada instancia de punto debe ser sometida a una tarea de desambiguación. Para esto utilizamos una herramienta externa, UDPipe (Straka et al., 2016), un analizador morfosintáctico que incluye esta función como parte del procesamiento del texto y con un desempeño aceptable en esta tarea específica. Para las mediciones de riqueza léxica de cada subcorpus utilizamos el cálculo del *type/token ratio* aplicando también para ello un script Perl, concretamente la estructura de *hash tables* que ofrece este lenguaje. Otros registros fueron la frecuencia de uso de marcadores discursivos, de deixis y de modalización, utilizando el software libre Text·a·Gram (Nazar, 2024), que permite detectar, marcar y clasificar estos rasgos discursivos en forma automática en grandes volúmenes de textos.

4. RESULTADOS Y DISCUSIÓN

4.1 Análisis de longitud de párrafos y oraciones

Los resultados de la longitud de párrafos y oraciones, medida en caracteres, se visualizan en diagramas de caja en las figuras 1 y 2. Estos resultados consideran la totalidad de los párrafos de licenciatura y de doctorado. Al comparar la extensión de los párrafos (figura 1) y la extensión de las oraciones (figura 2), se puede observar que los textos naturales presentan una mayor variabilidad, con un rango más amplio de dispersión. En contraste, los textos sintéticos muestran una menor dispersión y una distribución más acotada, es decir, una mayor homogeneidad estructural.

Por su parte, las figuras 3 y 4 muestran la distribución de la extensión de los párrafos naturales y sintéticos respectivamente, representada por histogramas que ilustran la cantidad de caracteres por párrafo. De forma similar, las figuras 5 y 6 presentan histogramas que muestran la extensión en caracteres de las oraciones para textos naturales y sintéticos.

[5] https://www.perl.org/

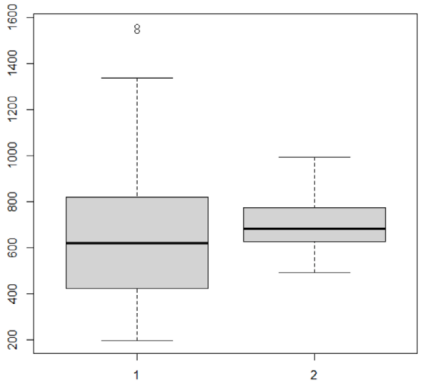

Figura 1. Extensión de los párrafos (en caracteres) en texto natural (1) y texto sintético (2).

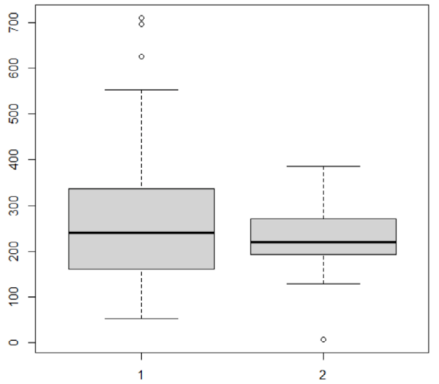

Figura 2. Extensión de las oraciones (caracteres) en texto natural (1) y texto sintético (2).

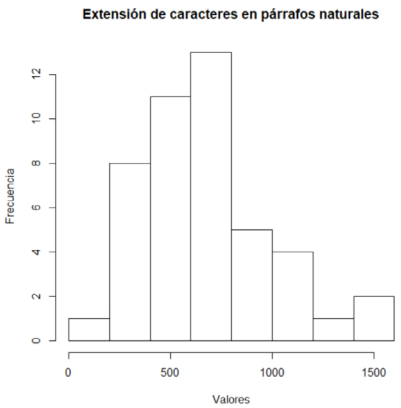

Figura 3. Extensión de párrafos naturales.

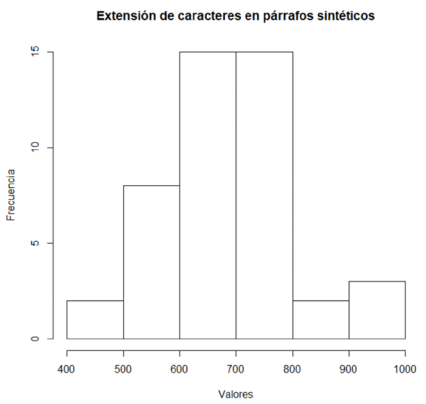

Figura 4. Extensión de párrafos sintéticos.

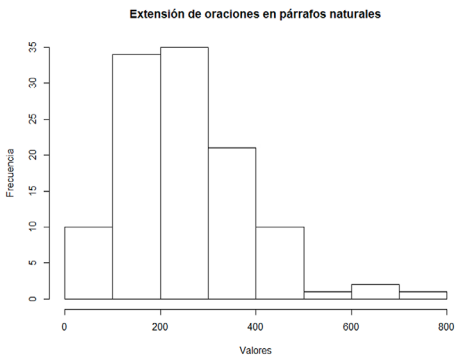

Figura 5. Extensión de oraciones naturales.

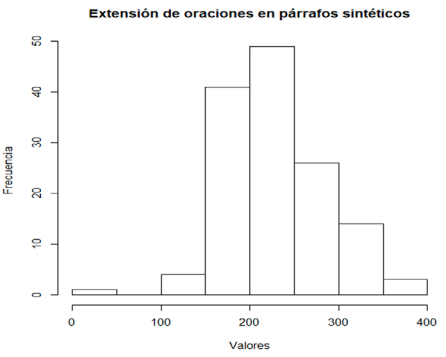

Figura 6. Extensión de oraciones sintéticas.

En el caso de los textos naturales, tanto la extensión de los párrafos como de las oraciones presentan una mayor variabilidad en su extensión, y la forma de distribución en el histograma indica que no siguen una distribución normal sino con un sesgo positivo. Esta distribución sesgada de la extensión de enunciados es un fenómeno ampliamente observado en textos naturales desde los primeros tiempos de la lingüística de corpus (Kučera & Francis, 1967). Lo que ahora se observa, en cambio, en el caso de los textos sintéticos, es que muestran menor variedad y una extensión más corta tanto en párrafos como en oraciones. Además, en este caso, la distribución del histograma parece acercarse más a una forma simétrica, aunque sin llegar a perfilarse claramente como una distribución normal.

Si bien no existen diferencias estadísticamente significativas entre ambas muestras según las pruebas basadas en la diferencia de medias (el *t-test* arroja p-value=0.6735 y p-value=0.0589, para los datos de párrafo y de oración, respectivamente, mientras la U-test de Mann-Whitney resulta en p-value = 0.43 para párrafo y p-value = 0.1386 para oración), sí es posible apreciar que los textos sintéticos son más homogéneos que los naturales, con una estructura más uniforme y consistente. La escritura natural, en cambio, presenta mayor diversidad, lo que se refleja en una mayor heterogeneidad en la longitud de los párrafos y oraciones.

4.2 Análisis riqueza léxica

A continuación, se presentan los resultados obtenidos de las diferentes medidas. En primer lugar, la tabla III muestra el resultado del cálculo de TTR. Aquí, en ambos casos (licenciatura y doctorado), en los subcorpus naturales se observa un valor mayor por margen de 7 puntos porcentuales.

Tabla III. *Type-Token Ratio* (TTR) del corpus.

	Lic. Natural	**Lic. Sintético**	**Doct. Natural**	**Doct. Sintético**	**Total Natural**	**Total Sintético**
Types	1167	968	686	621	1613	1296
Tokens	3031	2987	1471	1565	4502	4552
TTR	**0,39**	0,32	**0,47**	0,40	**0,36**	0,28

4.3 Frecuencia total de las variables analizadas

En cuanto a las mediciones de las variables discursivas (VD), los resultados obtenidos se presentan en las figuras 7, 8 y 9. Respectivamente, marcadores discursivos (MD), deixis (DX) y modalizadores (MO). Al igual que en el cálculo

de la riqueza léxica, los resultados aquí obtenidos de todas las variables discursivas indican una mayor frecuencia de uso en los textos naturales respecto a los sintéticos. En esta primera visión de conjunto, se aprecia que los textos naturales en todos los casos superan en frecuencia de uso de los diferentes mecanismos a los textos naturales, aunque la diferencia no llega a ser estadísticamente significativa (el test del chi cuadrado resulta en un p-value = 0,24).

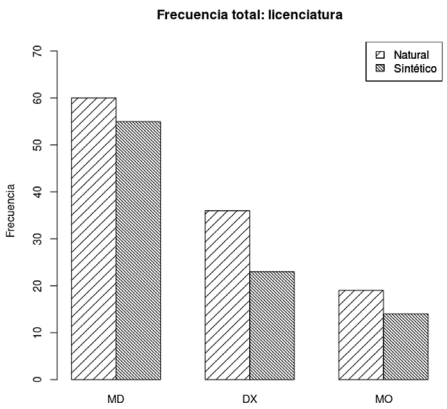

Figura 7. Frecuencias en licenciatura.

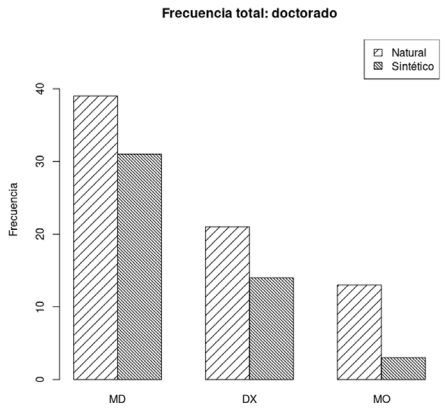

Figura 8. Frecuencias en doctorado.

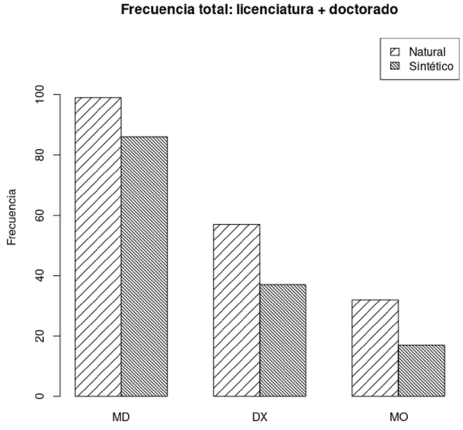

Figura 9. Frecuencia de VD en tesis de licenciatura y doctorado.

4.4 Análisis de marcadores discursivos

Los resultados de frecuencia de las distintas funciones de marcadores discursivos en textos naturales y sintéticos se exponen en las figuras 10 para licenciatura, 11 para doctorado y 12 para el total. Estos resultados presentan diferencias en licenciatura

y doctorado. En la licenciatura, los textos naturales utilizan más conectores consecutivos, aditivos, contraargumentativos y de refuerzo argumentativo. Los textos sintéticos, en cambio, tienen una tendencia a utilizar marcadores discursivos estructuradores, en particular el comentador *en este contexto*, además aditivos y de concreción. En el caso de doctorado, en los textos naturales el uso de marcadores discursivos está más distribuido y en la suma se puede apreciar que los textos naturales tienen un uso equilibrado de los marcadores pero con énfasis en los argumentativos, mientras que los sintéticos utilizan marcadores más neutros como los organizadores y aditivos. En este caso, el chi cuadrado sí arroja una diferencia estadísticamente significativa (p-value = 0.002).

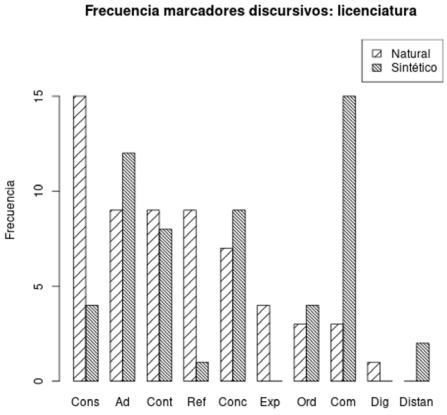

Figura 10. Frecuencias MD en licenciatura.

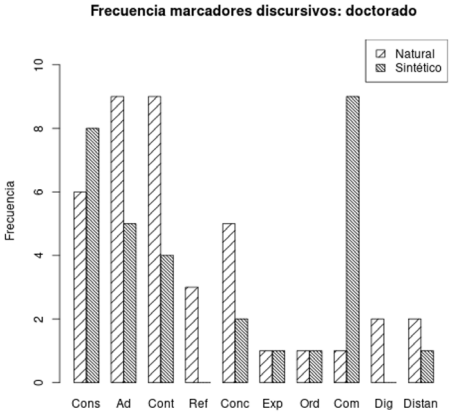

Figura 11. Frecuencias MD en doctorado.

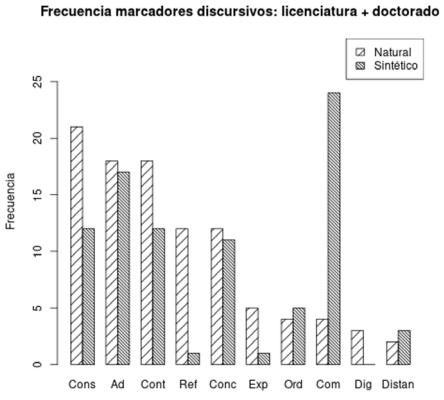

Figura 12. Frecuencias MD en tesis de licenciatura y doctorado.

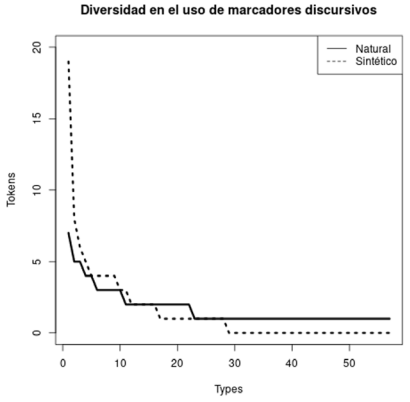

Figura 13. Diversidad de MD en texto natural vs. sintético.

La figura 13, en tanto, expone la diferencia en diversidad en el uso de los marcadores discursivos entre textos naturales y sintéticos, sumando licenciatura y doctorado. Se visualiza que los textos naturales poseen mayor diversidad en el uso de marcadores discursivos en comparación con los sintéticos. Los naturales tienen una variedad de 57 *types* sobre 99 *tokens* (0.57), mientras que los sintéticos poseen 28 *types* sobre 86 *tokens* (0.32), una diferencia mayor que la de TTR de léxico general. Cabe mencionar que de los 86 *tokens* del texto sintético, 19 corresponden a un solo *type,* el ya mencionado marcador textual *En este contexto*.

4.5 Análisis de deixis

Los resultados de la medición de deixis se muestran en la figura 14 para licenciatura, 15 para doctorado y 16 para la suma de ambas. Las gráficas se desglosan en las categorías de deixis temporal, espacial y actancial. En el caso de la deixis, las mediciones muestran menos diferencias, aunque destaca de cualquier modo la obtenida en el caso de la deixis actancial, indicio de una mayor presencia de marcas de subjetividad en el texto natural. En cualquier caso, las diferencias no llegan a ser lo suficientemente grandes como para alcanzar la significación estadística según el chi cuadrado (p-value = 0.15).

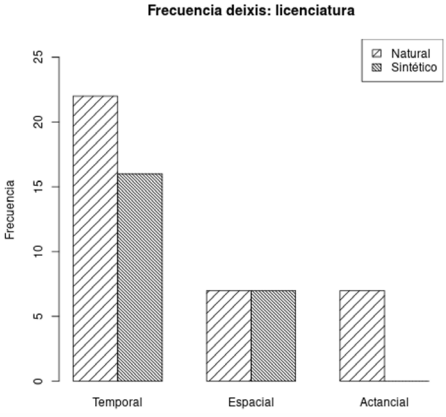

Figura 14. Frecuencias DX en licenciatura.

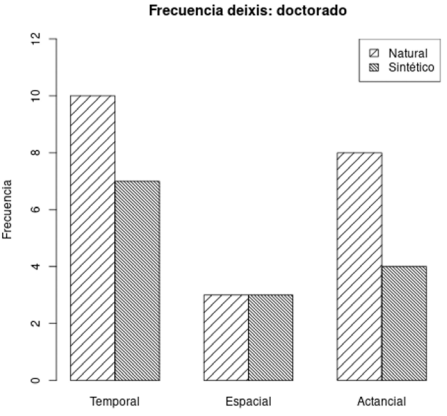

Figura 15. Frecuencias DX en doctorado.

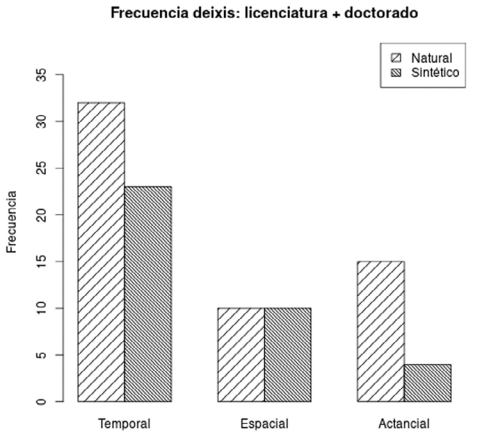

Figura 16. Frecuencia DX en tesis de licenciatura y doctorado.

4.6 Análisis de modalizadores

Los resultados de la medición de modalizadores en textos naturales y textos sintéticos se exponen en las figuras 17 para licenciatura, 18 para doctorado y 19 para el total. Las gráficas desglosan los modalizadores en las categorías de deónticos, aléticos, epistémicos, axiológicos y veredictorios.

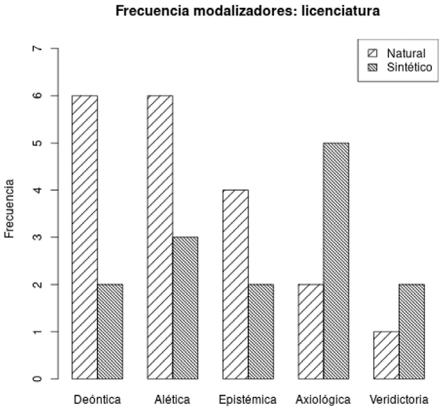

Figura 17. Frecuencias MO en licenciatura.

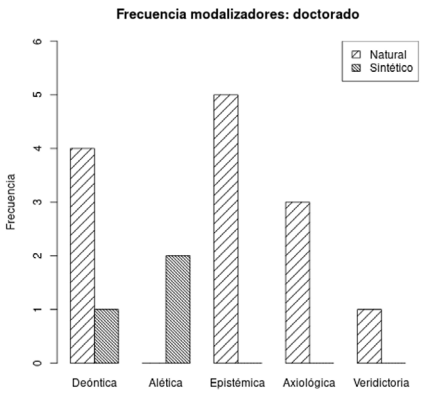

Figura 18. Frecuencias MO en doctorado.

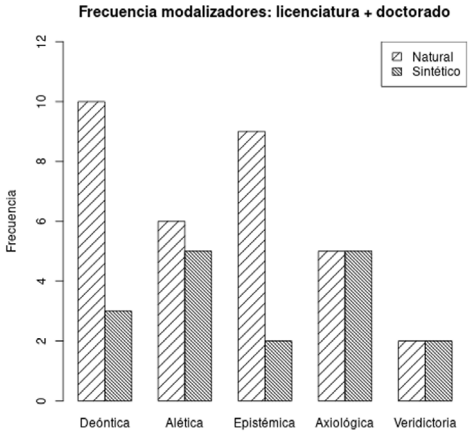

Figura 19. Frecuencia MO en tesis de licenciatura y doctorado.

Como en el caso de la deixis, en la modalización no se aprecian grandes diferencias entre textos naturales y sintéticos cuando se ve licenciatura y doctorado en conjunto, aunque sí se aprecia mayor uso de modalizadores deónticos y epistémicos en el caso de los textos naturales, es decir, nuevamente, marcas de la subjetividad que están menos presentes en el texto sintético. Nuevamente, sin embargo, los números son todavía demasiado bajos como para alcanzar una diferencia estadística (el chi cuadrado arroja un p-value = 0.39).

En cuanto a la diferencia entre licenciatura y doctorado, en el texto natural se utiliza en mayor medida los modalizadores deónticos y aléticos, mientras que en doctorado aumentan los epistémicos. Es interesante observar, además, que las lecturas del texto sintético también cambian según se trate de licenciatura y doctorado. Hay mayor uso de modalizadores axiológicos en el texto sintético de las tesis de licenciatura en comparación con el sintético de doctorado, lo que podría ser resultado de un efecto de imitación del ejemplo de párrafo natural utilizado en la instrucción dada en cada caso al chatbot.

5. CONCLUSIONES Y TRABAJO FUTURO

En este artículo hemos dado a conocer algunas diferencias observables, en cuanto a variables discursivas, entre textos naturales y textos sintéticos en el caso de las tesis de licenciatura y doctorado. Las mediciones de longitud oracional y de párrafo muestran un patrón característico, menos uniforme que el sintético. Asimismo, los textos naturales muestran mayor riqueza léxica y en general mayor frecuencia

y diversidad en el uso marcadores discursivos, deixis personal y modalizadores respecto a los textos sintéticos.

Entre los hallazgos más interesantes, se observó un uso frecuente de marcadores discursivos estructuradores (en particular, comentadores), y concretamente uno en específico, el marcador *en este contexto*, el cual parece una suerte de "marca personal" de ChatGPT. Además, la diferencia respecto a los tipos de marcadores utilizados por los dos tipos de texto, los textos naturales utilizaron marcadores discursivos típicos del género académico de argumentación propio de la tesis, mientras que los sintéticos mostraron un uso más genérico de marcadores como los estructuradores y aditivos. En cuanto a deixis y modalización, el texto natural concentra las que corresponden a marcas de subjetividad, que en cambio son poco frecuentes en el texto sintético.

Esta ha sido una primera exploración del tema y todavía presenta limitaciones. Varias de las pruebas realizadas no arrojaron diferencias estadísticamente significativas entre los dos grupos, probablemente por falta de potencia estadística al ser una muestra de tamaño insuficiente. Se abren, a partir de aquí, múltiples opciones de investigación. En este momento nos encontramos reproduciendo estos experimentos con muestras de mayor tamaño y de otros tipos de texto. Uno de ellos es un conjunto de 5000 artículos de investigación en lingüística y otro de 4000 columnas de opinión de diferentes periódicos. Estamos probando, además, con otros chatbots, en particular, los que utilizan LLM libres. El sistema Ollama[6] merece una mención especial, ya que es un chatbot que puede descargarse y ejecutarse en local, de manera programática, y sin ningún tipo de límite más allá del alto consumo energético que estas herramientas implican. En cualquier caso, la automatización total del proceso con este sistema permitirá alcanzar la potencia estadística necesaria para este tipo de estudios.

Finalmente, intentaremos, a partir de los resultados obtenidos, implementar un primer prototipo de clasificador de texto, en primera instancia, solo en castellano, en las categorías de natural y sintético. De tener éxito, el experimento podría tener un potencial interesante para las actuales prácticas académicas.

REFERENCIAS

Baayen, R. H (2008). *Analyzing linguistic data: a practical introduction to statistics using R*. Cambridge University Press.

Baker, T., & Smith, L. (2019). *Educ-AI-tion rebooted? Exploring the future of artificial intelligence in schools and colleges.* Retrieved from Nesta Foundation website:

[6] https://ollama.com/

https://media.nesta.org.uk/documents/Future_of_AI_and_education_v5_WEB.pdf

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* En Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). ACM. https://doi.org/10.1145/3442188.3445922

Borger, J. G., Ng, A. P., Anderton, H., Ashdown, G. W., Auld, M., Blewitt, M. E., Brown, D. V., Call, M. J., Collins, P., Freytag, S., Harrison, L. C., Hesping, E., Hoysted, J., Johnston, A., McInneny, A., Tang, P., Whitehead, L., Jex, A., & Naik, S. H. (2023). Artificial intelligence takes center stage: exploring the capabilities and implications of ChatGPT and other AI-assisted technologies in scientific research and education. *Immunology & Cell Biology*, 101(10), 923-935. https://doi.org/10.1111/imcb.12689

Briz, A., Pons, S., & Portolés, J. (2008). *Diccionario de partículas discursivas del español* [en línea]. Disponible en: http://www.dpde.es

Brown R. (1973). *A first language: the early stages*. Harvard University Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language models are few-shot learners. Advances in Neural Information Processing Systems*, 33, 1877-1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Calsamiglia, H., & Tusón, A. (1999). *Las cosas del decir. Manual de análisis del discurso*. Ariel.

Carlino, P. (2003). "Alfabetización académica: Un cambio necesario, algunas alternativas posibles", *Educere, Revista Venezolana de Educación,* vol. 6, núm. 20, pp. 409- 420 (en línea). Disponible en http://www.saber.ula.ve/bitstream/123456789/19736/1/ articulo7.pdf

Carlino, P. (2013). Alfabetización académica diez años después. *Revista mexicana de investigación educativa*, 18(57), 355-381

Casado Velarde, M. (1993). *Introducción a la gramática del texto del español.* Arco/Libros.

Cortés Rodríguez, L.M. (2001). Conectores, marcadores y organizadores como elementos del discurso. En J. J. de Bustos Tovar (Coord.), *Lengua, discurso, texto. I Simposio internacional de análisis del discurso,* vol. I (pp. 539-550). Madrid: Visor.

Cortés Rodríguez, L.M. & Camacho, M. M. (2005). *Unidades de segmentación y marcadores del discurso: elementos esenciales en el procesamiento discursivo oral.* Arco/Libros.

Crompton, H. & Burke, D. (2023) Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higer Education* 20 (22), https://doi.org/10.1186/s41239-023-00392-8

Cuenca, M. (2010). *Gramática del texto.* Arco Libros.

Desaire, H., Chua, A., Kim, M. & Hua, D. (2023) Accurately detecting AI text when ChatGPT is told to write like a chemist. *Cell Reports Physical Science* 4 (11) https://doi.org/10.1016/j.xcrp.2023.101672

Flesch, R. (1949). *The Art of Readable Writing*. Harper.

Fuentes Rodríguez, C. (2003). Operador/conector, un criterio para la sintaxis discursiva. Rilce. *Revista de Filología Hispánica*, 19(1), 61-85.

Fuentes Rodríguez, C. (2009). *Diccionario de conectores y operadores del español*. Arco/Libros.

García-Negroni, M., & Tordesillas, M. (2001). *La enunciación en la lengua: de la deixis a la polifonía*. Gredos.

García-Peñalvo, F. J., Llorens-Largo, F., & Vidal, J. (2024). La nueva realidad de la educación ante los avances de la inteligencia artificial generativa. *RIED: Revista Iberoamericana de Educación a Distancia*, 27(1), 9-39.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., Jansen, A., Gazula, H., Choe, G., Rao, A., Kim, C., Casto, C., Fanda, L., Doyle, W., Friedman, D., … Hasson, U. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3), 369-380. https://doi.org/10.1038/s41593-022-01026-4

González, C. & Ibáñez, R. (2017). Leer y escribir en contextos académicos. En R. Ibáñez & C. González (Eds.) *Alfabetización Disciplinar en la Formación Inicial Docente: Leer y escribir para aprender* (pp. 27-41). Ediciones Universitarias de Valparaíso.

Graesser, A., McNamara, D., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193-202.

Greimas, A.J. & Courtés, J. (1991). *Semiótica: Diccionario razonado de la teoría del lenguaje.* Gredos.

Herdan, G. (1964). *Quantitative linguistics*. Butterworths.

Holmes, W., & Tuomi, I. (2022). State of the Art and Practice in AI in Education. *European Journal of Education*, 57(4), 542-570. https://doi.org/10.1111/ejed.12533

Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100001. https://doi.org/10.1016/j.caeai.2020.100001

Katinskaia, A., & Yangarber, R. (2024). *GPT-3.5 for grammatical error correction.* arXiv. https://arxiv.org/abs/2405.08469

Kelih, E., Grzybek, P., Antić, G., & Stadlober, E. (2006). Quantitative Text Typology: The Impact of Sentence Length. En: M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, & W. Gaul(Eds) *From Data and Information Analysis to Knowledge Engineering. Studies in Classification, Data Analysis, and Knowledge Organization*. Springer.

Kerbrat-Orecchioni, C. (1997). La enunciación de la subjetividad en el lenguaje. Buenos Aires: Edicial.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). *Large language models are zero-shot reasoners.* arXiv. https://arxiv.org/abs/2205.11916

Kučera, H. & Francis, W. (1967). *Computational analysis of present-day American English*. Brown University Press.

Lobos, I. (en preparación). *Marcadores discursivos del género tesis y su distribución según movidas retóricas, nivel de inserción y disciplina.* Tesis doctoral. Pontificia Universidad Católica de Valparaíso.

Lozano, J., Peña-Marín, C., & Abril, G. (1989). *Análisis del discurso: hacia una semiótica de la interacción textual*. Cátedra.

Marinkovich, J., Sologuren, E., & Shawky, M. (2018). The process of academic literacy in Civil Engineering Computer Science. An approach to academic writing and its genres in a learning community. *Círculo de Lingüística Aplicada a la Comunicación*, 74, 195-220. https://doi.org/10.5209/CLAC.60520

Martín Zorraquino, M. A. & Portolés, J. (1999). Los marcadores del discurso. En I. Bosque y V. Demonte (eds.) *Gramática descriptiva de la lengua española*, vol. 2. Espasa, pp. 4051- 4213.

Meza, P., & Rivera, B. (2018). La comunicación del conocimiento propio en tesis: variación entre grados académicos en la sección desarrollo teórico. *RLA. Revista de Lingüística Teórica y Aplicada*, 56(1), 115-138. http://dx.doi.org/10.4067/S0718-48832018000100115

Montolío, E. (2001). *Conectores de la lengua escrita. Contraargumentativos, consecutivos, aditivos y organizadores de la información.* Ariel.

Morales-Chan, M.A. (2023). *Explorando el potencial de Chat GPT: Una clasificación de Prompts efectivos para la enseñanza.* Universidad Galileo (Repositorio Institucional). http://biblioteca.galileo.edu/tesario/handle/123456789/1348

Navarro, F. (2017). De la alfabetización académica a la alfabetización disciplinar. En R. Ibáñez y C. González (Ed.) *Alfabetización Disciplinar en la Formación Inicial Docente: Leer y escribir para aprender* (pp. 7-15). Ediciones Universitarias de Valparaíso.

Nazar, R. (2024). Statistical modeling of discourse genres: the case of the opinion column in Spanish. *SN Computer Science* 5(959):1-11.

Obreque, J., & Nazar, R. (2023). Detección de operadores modales: una primera exploración en castellano. *Linguamatica*. 15(2): 37-49. https://linguamatica.com/index.php/linguamatica/article/view/411

OpenAI. (2022, 30 de noviembre). *OpenAI introduces ChatGPT, a new AI-powered chatbot.* https://openai.com/blog/chatgpt

Parodi, G., Venegas, R., Ibáñez, R. & Gutiérrez, R. (2008). Géneros del discurso en el Corpus PUCV-2006: Criterios, definiciones y ejemplos. En Giovanni Parodi (editor), *Géneros académicos y géneros profesionales: Accesos discursivos para saber y hacer* (pp. 39-73). Valparaíso. Universitarias de Valparaíso

Pérez Canales, J. (2009) *Marcadores de modalidad epistémica: un estudio contrastivo (francés-español)*. Universitat de Valencia. Tesis Doctoral.

Popenici, S. A., & Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*.

Pyatkin, V., Sadde, S., Rubinstein, A., Portner, P., & Tsarfaty, R. (2021). *The possible, the plausible, and the desirable: Event-based modality detection for language processing.* ArXiv [cs.CL]. 10.48550/arXiv.2106.08037.

Rey, M., & Velásquez, E. (2023). La escritura de la tesis: concepciones, creencias y actitudes de doctorandos en educación. *Innovación educativa* 23 (92), 10-34.

Robledo, H. (2021). *Categorización de los marcadores del discurso del español: una propuesta inductiva guiada por corpus paralelo* (Tesis doctoral, Pontificia Universidad Católica de Valparaíso). Disponible en https://catalogo.pucv.cl/cgi-bin/koha/opac-detail.pl?biblionumber=434483

Robledo, H. & Nazar, R. (2023). A proposal for the inductive categorisation of parenthetical discourse markers in Spanish using parallel corpora. International *Journal of Corpus Linguistics*, 28(4): 500-527. https://doi.org/10.1075/ijcl.20017.rob

Rojo, G. (2021). *Introducción a la lingüística de corpus en español.* Routledge.

Russell, S., & Norvig, P. (2010). *Artificial intelligence - a modern approach.* Pearson Education.

Sabzalieva, E., & Valentini, A. (2023). ChatGPT e inteligencia artificial en la educación superior: guía de inicio rápido. Unesco. https://unesdoc.unesco.org/ark:/48223/pf0000385146_spa

Santos Río, L. (2003). *Diccionario de partículas.* Luso-Española de Ediciones.

Straka, M., Hajič, J., & Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) (pp. 4290-4297). *European Language Resources Association (ELRA)*. https://www.aclweb.org/anthology/L16-1680/

Van Dijk, T. (2012). *Discourse and Context: A Sociocognitive Approach.* Cambridge University Press.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, Ł. & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*. 30.

Velásquez, M., & Marinkovich, J. (2016). Hacia un modelo explicativo del proceso de alfabetización académica en las licenciaturas en Historia y Biología. *RLA. Revista de Lingüística Teórica y Aplicada,* 54(2), 113-136. https://doi.org/10.4067/S0718-48832016000200006

Venegas, R., Zamora, S. & Galdames, A. (2016). Hacia un modelo retórico-discursivo del macrogénero Trabajo Final de Grado en Licenciatura. *Revista Signos*. Estudios de Lingüística, 49(1), 247-279.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*. 16:39.