ARTÍCULOS

Círculo de Lingüística Aplicada a la Comunicación
ISSN-e: 1576-4737

EDICIONES COMPLUTENSE

https://dx.doi.org/10.5209/clac.79977

Análisis automático de la complejidad sintáctica de textos escolares

Romualdo Ibáñez Orellana[1]; Juan Zamora Osorio[2]; Mariela Cisnero Correa[3]; Solange Aguirre Rozas[4]

Recibido: 21 de enero de 2022 / Aceptado: 16 de abril de 2022

Resumen. El objetivo del presente estudio es comparar la complejidad sintáctica de los textos utilizados en tres asignaturas de educación primaria. Para ello, se recolectó un corpus compuesto por 2121 textos, presentes en los textos escolares entregados por el Estado de Chile al estudiantado de colegios públicos. Se realizó un análisis automático, por medio de un algoritmo que identifica las relaciones de dependencia sintáctica entre los constituyentes de una oración y, posteriormente, calcula la Longitud de Dependencia Sintáctica (LDS) promedio de la misma. Los resultados revelaron que la LDS promedio de los textos analizados, correspondientes a diferentes cursos y asignaturas, es homogéneamente baja. Del mismo modo, se observó que no existe un patrón de complejización a medida que avanzan los cursos. También quedó en evidencia que, si bien no fue posible apreciar patrones disciplinares que permitieran determinar la existencia de asignaturas con mayor CS, sí existe una tendencia que sitúa a Historia, Geografía y Ciencias Sociales como la más compleja de las asignaturas analizadas, en términos de configuración sintáctica.
Palabras clave: Complejidad Sintáctica; Longitud de Dependencia Sintáctica; Textos Escolares; Análisis Automático.

[en] Automatic Analysis of school textbooks' syntactic complexity

Abstract. The purpose of this study was to compare the syntactic complexity of texts used to communicate knowledge in the school textbooks of three school subjects. To do so, we collected a corpus of 2121 texts, used in the school textbooks that the State of Chile provides to students attending public schools. Texts were automatically analyzed by an algorithm that identifies syntactic dependency relations in a sentence and then calculates the mean Syntactic Dependency Length (SDL) of that sentence. Results showed that the SDL of the analyzed texts -corresponding to different levels and school subjects- was homogeneously low. Besides, it was possible to observe that there was not a pattern of incremental complexity associated with school levels. Results also showed that while it was not possible to identify disciplinary patterns that allowed the identification of school subjects exhibiting more CS, there was a tendency that places History, Geography and Social Science as the most syntactically complex.
Keywords: Syntactic Complexity; Syntactic Dependency Length; School textbooks; automatic analysis.

Cómo citar: Ibáñez Orellana, R., Zamora Osorio, J., Cisnero Correa, M., & Aguirre Rozas, S. (2022). Análisis automático de la complejidad sintáctica de textos escolares. *Círculo de Lingüística Aplicada a la Comunicación* 91, 127-142.

[1] Pontificia Universidad Católica de Valparaíso.
Correo electrónico: romualdo.ibanez@pucv.cl
ORCID: https://orcid.org/0000-0001-9298-3806

[2] Pontificia Universidad Católica de Valparaíso.
Correo electrónico: juan.zamora@pucv.cl
ORCID: https://orcid.org/0000-0003-0003-182X

[3] Pontificia Universidad Católica de Valparaíso.
Correo electrónico: mariela.cisnero.c@mail.pucv.cl
ORCID: https://orcid.org/0000-0002-5747-5475

[4] Pontificia Universidad Católica de Valparaíso.
Correo electrónico: solange.aguirre.r@mail.pucv.cl
ORCID: https://orcid.org/0000-0003-0828-7133)

1. Introducción

Como herramienta pedagógica de uso frecuente, el texto escolar (TE) cumple un rol fundamental en los procesos de enseñanza-aprendizaje, razón por la que ha sido ampliamente estudiado durante las últimas décadas (Möenne & López, 2007; Bañados, 2007; Olivares, 2007; Oteíza, 2009; Altamirano *et al.*, 2014). Entre la gran variedad de investigaciones existentes sobre el TE, resultan de interés aquellas que –dada la indiscutible influencia de las características de los textos en su procesamiento y comprensión– han puesto atención a sus rasgos lingüísticos específicos (Kasule, 2011; McNamara *et al.*, 2014). En Chile, algunas investigaciones de este tipo han permitido observar que tanto los géneros discursivos utilizados para comunicar el conocimiento en el TE (Ibáñez *et al.*, 2017), como las funciones que estos desempeñan (Ibáñez *et al.*, 2018) varían dependiendo de la asignatura. Otras más recientes han comenzado incluso a indagar en los rasgos lingüísticos del TE, haciendo uso de herramientas informáticas. Entre estos estudios, se destaca el realizado por Rojas *et al.* (2020), quienes realizan un análisis semiautomático de los rasgos lingüísticos de TE de Lenguaje y Comunicación (LC), con el propósito de determinar su lecturabilidad. Los resultados de este trabajo demostraron que la mayor parte de los textos analizados son muy complejos para sus destinatarios, y que su dificultad se evidencia principalmente en el nivel oracional.

Estos hallazgos indican que la complejidad sintáctica (CS) podría constituir un rasgo fundamental para determinar la dificultad de procesamiento de los textos (Poulsen & Gravgaard, 2016; Kleijn, 2018). Sin embargo, aún no se cuenta con datos empíricos que permitan conocer el nivel de CS de los textos utilizados para comunicar el conocimiento en los TE chilenos, tampoco se sabe si ese nivel varía entre asignaturas o si aumenta con el nivel de escolaridad. Pero, tal vez, más importante aún, no existe evidencia de que tal cálculo se haya realizado de manera automática. Por tal razón, el objetivo del presente estudio es comparar, por medio de una herramienta de análisis automático, la complejidad sintáctica de los textos utilizados para comunicar el conocimiento en textos escolares de tres asignaturas de educación básica. Este trabajo contribuye a las investigaciones sobre la CS del TE en dos direcciones complementarias. Por un lado, compara la CS de 2121 textos extraídos de TE de tres asignaturas de educación básica. Por otro lado, representa un esfuerzo concreto por adaptar un método de cálculo automático de promedio de Longitud de Dependencia Sintáctica (LDS) para el español. A continuación, se abordan los supuestos teóricos que sustentan esta investigación, luego, se da cuenta de los aspectos metodológicos. Posteriormente, se consignan y discuten los resultados, con particular énfasis en las eventuales implicancias pedagógicas. Por último, se ofrecen las conclusiones.

2. El texto escolar y la variación disciplinar

El TE constituye uno de los recursos didácticos más importantes para llevar los contenidos del currículum al aula. Además de presentar métodos de aprendizaje mediante los cuales se favorece la adquisición de nuevos conocimientos y habilidades disciplinares, contribuye exponiendo valores y normas que los usuarios interiorizan para orientar su propio desarrollo personal (Ramírez, 2002; Choppin, 2000). Desde una perspectiva discursiva (Christie, 2002), el TE es asumido como un macrogénero, cuyo macropropósito comunicativo es instruir respecto de conocimiento procedural y declarativo (Martin & Rose, 2008; Rose, 2014; Ibáñez *et al.*, 2019). La investigación en torno a este macrogénero ha demostrado que está constituido por dos tipos de géneros: los géneros del conocimiento (GCO), a través de los cuales el conocimiento es presentado (*Definición*, *Diccionario*, *Mapa*, etc.) y los géneros curriculares (GCU), los cuales permiten orientar y dirigir la actividad pedagógica (*índice*, *planificación*, *ejercicios*, etc.) (Christie, 1998; Rose, 2014).

Según Martin y Rose (2013), cada disciplina utiliza diferentes GCO para comunicar el conocimiento, fenómeno asociado con la recontextualización del conocimiento (Christie, 2002) y observable en los contextos pedagógicos y especialmente en el TE. Con el propósito de indagar en tal variación, Ibáñez *et al.* (2017) analizaron un corpus compuesto por 100 textos escolares de las asignaturas de *Matemáticas* (M), *Lenguaje y Comunicación* (LC), *Historia, Geografía y Ciencias Sociales* (HGCS), *Inglés* (I) y *Ciencias Naturales* (CN). Identificaron 15 GCO (*Definición, Biografía, Nota, Guía Procedural, Exposición de Contenido*, etc.), a partir de los cuales pudieron clasificar 11139 instancias textuales. Otro hallazgo interesante del estudio fue que los GCO predominantes en cada asignatura eran diferentes. En HGCS, el predominante fue *Fuente Histórica*; en M, *Guía Procedural*; en LC, *Definición*; y en CN, *Exposición de Contenido*. También se observó que había géneros que ocurrían solo en una asignatura: *Fuente Histórica* ocurría solo en HGCS, la *Noticia de Divulgación Científica*, solo en CN y *Expresiones Frecuentes*, solo en I.

Entre los GCO identificados, resultó particularmente interesante el género *Exposición de Contenidos* (EC), ya que fue uno de los más frecuentes (22.9 %) en relación con las ocurrencias totales. Este género es conceptualizado como un género pedagógico cuyo propósito comunicativo es presentar temáticas propias de una asignatura de acuerdo con una organización discursiva expositiva o narrativa (Ibáñez *et al.,* 2017). Generalmente, en el TE cumple la función didáctica de exponer los contenidos centrales de una unidad o lección (Ibáñez *et al.,* 2018). En el Anexo 1, se observa un ejemplo extraído de un texto escolar de la asignatura LC de octavo año básico.

La variación observada en los hallazgos anteriormente expuestos puede ser explicada por medio del supuesto de que, según su naturaleza, las distintas disciplinas poseen maneras diferentes para construir y comunicar el conocimiento (Hyland, 2004; Wheelahan, 2010), lo que sería perceptible en distintos rasgos lingüísticos y discursivos. Asumiendo tal supuesto, Ibáñez *et al.,* (2019) compararon el uso de relaciones de coherencia en los TE de cuatro asignaturas de educación básica. Los resultados demostraron que mientras algunas relaciones eran utilizadas en las cuatro asignaturas (*Conjunción*, *Descripción de Conceptos*), otras eran utilizadas casi exclusivamente en algunas de ellas (*Condición Evento* en CN, *Contraste Básico* en HGCS, *Deíctica* en LC y *Condición Pregunta* en M). Resultados similares obtuvieron Santana *et al.* (2021) al comparar el uso de expresiones conectivas utilizadas para marcar relaciones causales en TE de cuatro asignaturas. Identificaron 76 mecanismos lingüísticos, de los cuales la mayoría desempeñaba un rol polifuncional, mientras que, en un número menor, observaron patrones de especificidad por asignatura.

En cuanto a las características sintácticas, Bailey *et al.* (2007) compararon las selecciones gramaticales en un corpus compuesto por 12 textos de quinto grado de escuela básica en tres asignaturas: M, CN y Estudios Sociales. Los resultados también revelaron diferencias disciplinares, las cuales se observaron en el mayor uso de oraciones simples en los textos de M que en los de CN y Estudios Sociales y, en correspondencia, un mayor uso de oraciones complejas en las últimas dos asignaturas que en M. En contraste con estos resultados, Graesser *et al.* (2011), empleando la herramienta Coh-Metrix (Graesser & McNamara, 2011), realizaron un análisis automático de un corpus de más de 37000 textos que abarcaban desde el kindergarten hasta el 12º grado de escolaridad. Estos textos fueron clasificados en tres áreas temáticas: Lenguaje-Artes, Ciencias y Estudios Sociales-Historia. Los autores hallaron que los textos de Lenguaje-Artes son más complejos sintácticamente, seguidos por los textos de Estudios Sociales-Historia y los de Ciencias.

3. Estándares del texto escolar según el Estado de Chile

El currículum de Educación Básica del Estado de Chile contempla la formación de lectores eficientes, capaces de extraer y construir el significado de los textos escritos, no solo a nivel literal sino a nivel interpretativo. Según MINEDUC Chile (2015, 2018a, 2018b), la comprensión de un texto implica la extracción de información, hacer inferencias relacionadas con los aspectos no dichos explícitamente y hacer evaluaciones críticas acerca de los contenidos que lee. Esto supone que el lector tiene un papel activo, en el que relaciona sus conocimientos previos con la información que se extrae del texto.

Es prioridad de la escuela básica la formación de lectores activos y críticos, que acudan a la lectura como medio de información, aprendizaje y recreación en distintos ámbitos de la vida. La formación de estos individuos competentes contempla el carácter progresivo del aprendizaje (Gysling & Meckes, 2011), concibiéndolo como un continuo que se enriquece a lo largo de la trayectoria escolar. Desde esta perspectiva, la selección de lecturas destinadas a evaluar la comprensión está basada en la edad y etapa de desarrollo de las y los estudiantes, entendiéndose que, a medida que avanzan en el grado de escolaridad, los textos utilizados aumentan su grado de dificultad en base a criterios relacionados con el contenido y la forma (MINEDUC Chile, 2015). Tomando en cuenta estos aspectos, se busca que las lecturas utilizadas en los distintos niveles presenten un equilibrio entre ser comprensibles, para que las y los estudiantes se consideren competentes frente a la tarea, y a la vez lo suficientemente desafiantes para que las y los estudiantes progresen en sus habilidades de lectura y se sientan estimulados (MINEDUC Chile, 2015, 2018b).

Dado que las lecturas son la materia prima para lograr los objetivos, el estado contempla criterios que contribuyan con la selección de TE. Uno de estos criterios es precisamente la complejidad sintáctica, la cual está relacionada con los aspectos gramaticales. Los textos de educación básica presentan, en palabras del MINEDUC Chile (2015, 2018a), una sintaxis relativamente compleja, que se va incrementando a medida que aumenta el nivel de escolaridad. Se entiende la complejidad sintáctica del texto como presencia de cláusulas subordinadas que incrementan la dificultad de las estructuras oracionales empleadas para presentar los temas.

Considerando los hallazgos anteriormente expuestos, así como la relevancia de la CS en el procesamiento de los textos, resulta interesante indagar en la forma en que su uso podría variar dependiendo de las asignaturas y los niveles de escolaridad. En el apartado siguiente, se ahonda en este rasgo particular.

4. Complejidad sintáctica y la Longitud de Dependencia Sintáctica

La CS es entendida como el grado de dificultad que podría representar una construcción sintáctica a partir de sus constituyentes y a las relaciones que se establecen entre los mismos. Según Kleijn (2018), la CS determina la dificultad o facilidad con la que una estructura oracional es procesada y comprendida. Cuando una oración compleja es parte de un texto, el decrecimiento en la comprensión puede extenderse incluso más allá de tal oración y comprometer la calidad de la representación que se está generando (Gernsbacher, 1989). En este sentido, si la representación es incoherente o incompleta, la integración de nuevas oraciones también se verá

comprometida, lo cual afectará a la construcción del modelo de situación (Poulsen & Gravgaard, 2016; Kleijn, 2018).

La noción de CS fue inicialmente propuesta por Kellog Hunt (1965, 1970), quien, adhiriendo a los supuestos de la Gramática Generativa Transformacional, establece índices para la operacionalización del constructo por medio de lo que denominó la Unidad T o Unidad Mínima Terminal. Por medio de esta propuesta, Hunt (1970) demostró que, en inglés, la CS aumenta a medida que aumenta la edad y el nivel de escolaridad, lo cual se refleja en un aumento en la aparición de cláusulas subordinadas en las producciones lingüísticas de los niños. Tal hallazgo fue replicado para el español por Véliz (1988, 1999). Sin embargo, la Unidad T no ha sido la única forma de operacionalizar la CS, puesto que el gran interés de los especialistas por el fenómeno (Arnold *et al.,* 2000; Fedorenko *et al.*, 2013; Levy & Keller, 2013; Frantz *et al.*, 2015; Poulsen & Gravgaard, 2016; Staub, 2010; Kleijn, 2018; Futrell *et al.*, 2020, entre otros), ha resultado en un gran número de estudios al respecto y, por la misma razón, en la existencia de distintas maneras de operacionalización, tales como, número de palabras en una oración (Hawkins, 1990; Arnold *et al.,* 2000), relaciones que se establecen entre categorías sintácticas (Gili Gaya, 1972; Sedano, 2011), cantidad de nodos que dominan una estructura determinada (Chomsky, 1970, 1978; Ferreira, 1991; Givón, 1991, 2009; Hawkins, 1994; Rickford *et al.,* 1995) o las relaciones entre cláusulas (Nir & Berman, 2010; Sedano, 2011) y entre grupos de cláusulas en el texto (Nir & Berman, 2010). Independiente de la forma en que la CS ha sido operacionalizada, diversas investigaciones han podido dar cuenta de la relación entre CS, edad y grado de escolaridad (Herrera Lima, 1991; Rodríguez Fonseca, 1991; Vásquez, 1991; Crespo *et al.*, 2013; Aravena & Hugo, 2016; Peñaloza *et al.*, 2017; Sánchez & De Mier, 2017; Bartolomé, 2021, entre otros).

En la presente investigación, la CS se operacionaliza tomando en consideración la Longitud de Dependencia Sintáctica (LDS) (Gibson 1998, 2000; Grodner & Gibson, 2005), es decir, la distancia que separa a los constituyentes nucleares y sus correspondientes elementos dependientes en una oración. Así, una estructura oracional es más compleja en la medida en que la LDS entre sus elementos constituyentes sea mayor, esto es, mientras existan más elementos lingüísticos que se interpongan entre los núcleos sintácticos y sus dependientes, como puede ser, por ejemplo, un verbo y un sintagma nominal en función de sujeto u objeto (Gibson 1998, 2000; Grodner & Gibson, 2005; Kleijn, 2018; Futrell *et al.,* 2020). Un ejemplo de esto lo vemos en (1) y (2).

(1) Al igual que otras civilizaciones del mundo antiguo, *Atenas* tuvo un gobierno monárquico.
(2) *Atenas*, al igual que otras civilizaciones del mundo antiguo, tuvo un gobierno monárquico.

Entre las oraciones (1) y (2) existe una diferencia en la distancia que separa el sujeto gramatical *Atenas* y el núcleo verbal *tuvo*. En el primer caso, el verbo se encuentra en una posición adyacente con relación al sujeto. Esta distancia, según el cálculo automático realizado en esta investigación, es 1. Por su parte, en (2), el verbo está separado del sujeto por una frase adverbial que contiene 8 palabras. En este caso, la distancia entre el núcleo y su correspondiente dependiente (LDS) se incrementa a 9, dando como resultado una oración sintácticamente más compleja.

5. La LDS y su incidencia en el procesamiento y comprensión de textos escritos

La LDS es un factor relevante en el procesamiento y comprensión de textos escritos, pues constituye un factor determinante en la velocidad y eficacia en los procesos de integración gramatical necesarios para construir una representación mental coherente de lo que se lee (Grodner & Gibson, 2005; Fedorenko *et al.,* 2013; Kleijn, 2018). Una noción esencial para comprender el efecto de la LDS en el procesamiento y comprensión de textos es el concepto de localidad (Gibson, 1998, 2000; Grodner & Gibson, 2005; Demberg & Keller, 2008; Bartek *et al.,* 2011; Vasishth & Drenhaus, 2011; Futrell *et al.,* 2020). Se puede decir que una relación sintáctica es local cuando el núcleo sintáctico y sus dependientes son adyacentes**,** mientras que una relación sintáctica es no local cuando núcleo y dependientes están separados por otros elementos (Kleijn, 2018).

Diversos estudios psicolingüísticos han puesto en evidencia que los costos de procesamiento de relaciones no locales comienzan a observarse con pocas palabras intercaladas entre los elementos oracionales que deben ser integrados, lo cual se explica en función de las limitaciones de la memoria de trabajo para procesar la distancia entre elementos lingüísticos sintácticamente dependientes cuando se interponen entre estos nuevos elementos lingüísticos (King & Just, 1991; Gibson, 1998, 2000; Lewis & Vasishth, 2005; Grodner & Gibson, 2005; Vasishth & Drenhaus, 2011; Fedorenko *et al.,* 2013). Así las cosas, a medida que aumenta la distancia entre un núcleo sintáctico y sus dependientes, los costos de procesamiento se incrementan, de modo que es más fácil procesar una dependencia cuando hay un solo elemento entre núcleo y dependiente, que cuando hay más (Kleijn, 2018). Sin embargo, aún no existe certeza entre los especialistas respecto del número exacto de elementos interpuestos que hace que una relación no local comience a volverse problemática en términos de procesamiento y comprensión.

Algunas investigaciones demuestran que la interposición de al menos tres palabras es suficiente para evidenciar los efectos de localidad. Un ejemplo de esto es la investigación de Grodner y Gibson (2005) quienes, usando como método la lectura a propio ritmo, hallaron un incremento en los tiempos de lectura en oraciones subordinadas en dos condiciones: i) interposición de una frase preposicional de tres palabras entre el verbo de la subordinada y su correspondiente sujeto, y ii) interposición de una cláusula relativa de cinco palabras entre el verbo de la subordinada y el sujeto. Resultados similares son reportados por Bartek *et al.* (2011), quienes replican el trabajo de Grodner y Gibson (2005), usando medidas de seguimiento ocular. Los autores comparan el efecto de localidad en el procesamiento de oraciones doblemente incrustadas y no incrustadas, en las cuales se incrementó la distancia entre verbo y sujeto con una frase preposicional (3 palabras) y una cláusula relativa (5 palabras). Entre los hallazgos más interesantes se encuentran la presencia de efectos de localidad en las oraciones más simples a partir de la introducción de la frase preposicional. Este resultado, observado en las medidas tempranas, constituye una evidencia de los incrementos en los tiempos de procesamiento, a partir de una base de tres palabras, en la primera lectura de la oración.

En consonancia con estos hallazgos, Nicemboin *et al.* (2015), reportan efectos de localidad mediante el incremento de la distancia entre elementos sintácticos dependientes, mediante la adición de una frase adverbial de tres palabras. Los resultados de la investigación, llevada a cabo con protocolos de lectura a propio ritmo y medidas de movimiento ocular, dieron cuenta de un aumento en los tiempos de lectura y fijaciones con la adición de este elemento. Finalmente, Kleijn (2018) muestra un incremento en el tiempo de procesamiento en las oraciones con relaciones no locales, con una media de cuatro palabras. Además de los hallazgos respecto del procesamiento, también se han observado efectos en la comprensión. En este sentido, evidencia empírica en estudios de comprensión realizados con el test de HyTec Cloze (Kleijn, 2018) demuestran que, efectivamente, el desempeño disminuye en las pruebas asociadas a textos con mayor longitud de dependencia sintáctica, tanto en textos de información pública (Kleijn, 2018), como en textos escolares de diferentes asignaturas (Rojas, 2021). Dados los resultados de las investigaciones reportadas, es evidente que la CS puede ser un factor relevante en el procesamiento, la comprensión y el aprendizaje a partir del TE (Goldman, 1997; McNamara, 2004; McNamara & Kintsch, 1996), por lo que resulta relevante conocer su CS y si esta varía dependiendo de las asignaturas.

6. Análisis automático de complejidad sintáctica mediante cálculo de LDS

Gran parte del análisis automático de la CS se ha sustentado en el análisis de dependencias. Este enfoque de análisis se basa en el supuesto de que la estructura sintáctica consiste en relaciones binarias asimétricas entre palabras, denominadas relaciones de dependencias (Tesnière, 2015). En la actualidad, el interés por desarrollar descripciones y clasificaciones más precisas ha llevado a especialistas en el área del procesamiento del lenguaje natural y de la informática a utilizar métodos tales como *Support Vector Machine* (SVM) (Yamada & Matsumoto, 2003) y Redes Neuronales Artificiales (Chen & Manning, 2014; Dyer *et al.*, 2015) en el análisis de relaciones de dependencia sintáctica.

La identificación de relaciones de dependencia mediante este tipo de métodos estadístico-computacionales utiliza un corpus anotado con la estructura sintáctica de las oraciones como modelo base para ajustar de manera óptima sus parámetros internos. Luego, el modelo ya ajustado es usado para identificar las relaciones entre palabras en nuevas oraciones. Existen dos grandes grupos de modelos en este enfoque (Nivre & McDonald, 2008; Kübler *et al.,* 2009): basados en grafos (Koo *et al.* 2008; Hall *et al.*, 2014) y basados en transiciones (Watanabe & Sumita, 2015; Kitaev & Klein, 2018). En el primero, se considera cada arco posible en una oración y establece una función de puntuación que permite establecer qué tan deseable es cada uno. De esta manera, cada árbol tendrá una puntuación consistente en la suma de las puntuaciones de sus arcos. Finalmente, el problema es equivalente a encontrar el árbol con mayor puntuación para la oración. El segundo enfoque, basado en transiciones, aborda el problema como una toma óptima de decisiones, en donde las palabras son leídas secuencialmente y combinadas incrementalmente en estructuras sintácticas. Para este último enfoque, el problema consiste en identificar qué decisión tomar en cada paso, considerando además todas las decisiones tomadas hasta ese momento. Este último grupo ha sido desarrollado exitosamente empleando redes neuronales recurrentes como mecanismo de decisión para el idioma chino (Li *et al.*, 2018), checo, inglés y español (Dyer *et al.,* 2015; Ballesteros *et al.*, 2016; Marcheggiani *et al.,* 2017).

Las relaciones de dependencia sintáctica han probado no solo ser una forma útil de operacionalizar la CS de los textos, sino también un indicador relevante de su lecturabilidad (Falkenjack *et al.,* 2013; Pilán *et al.*, 2014; Futrell *et al.*, 2015; Falkenjack *et al.*, 2016; Chatzipanagiotidis *et al.*, 2021). Por la misma razón, algunos autores las han utilizado para calcular automáticamente la CS mediante el cálculo de la LDS de las oraciones.

Una de esas propuestas es la fórmula propuesta por Oya (2011), quien se propuso caracterizar la complejidad de un texto mediante el promedio de las sumas de las distancias entre cada par de palabras núcleo y dependiente contabilizadas linealmente en cada oración. Para este fin, un texto es analizado creando una red dirigida acíclica (DAG) por oración, conectando cada núcleo y su dependiente mediante un arco. Finalmente, el valor

del indicador estará dado por el promedio de las sumas totales de distancias sobre todas las oraciones. El uso de la distancia de dependencias promedio para describir la CS de los textos ha mostrado ser un indicador muy útil. En este contexto, Liu (2008) observó que los textos escritos mantienen una distancia de dependencias promedio dentro de un umbral entre 1,8 y 3,7.

7. Metodología

Como ya se señaló, este estudio tiene por objetivo comparar, por medio de una herramienta de análisis automático, la CS de los textos utilizados para comunicar el conocimiento en textos escolares de tres asignaturas de educación básica. Con base en la literatura revisada y, en este sentido, reconociendo el fenómeno de la recontextualización del conocimiento (Christie, 2002; Ibáñez *et al.,* 2017; 2018), nuestra hipótesis es que el nivel de complejidad sintáctica variará entre las asignaturas. También, asumiendo los estándares propuestos por el Estado de Chile para la creación de TE, esperamos que la CS aumente a medida que aumenta el nivel escolar. En esta sección describimos los procedimientos metodológicos realizados para alcanzar nuestro objetivo.

7.1 Corpus

El corpus está constituido por 2121 textos de los niveles de sexto, séptimo y octavo básico, correspondientes al género Exposición de Contenido. Según los resultados de Ibáñez *et al*. (2017), este género es el más utilizado para comunicar el conocimiento en los TE de Lenguaje y Comunicación (LC), Ciencias Naturales (CN) e Historia y Geografía y Ciencias Sociales (HGCS). Los TE corresponden a las versiones entregadas de manera gratuita por el Estado Chileno a las y los estudiantes de Educación Básica, entre los años 2012 y 2019 y producidos por las editoriales Cal y Canto, Galileo, Santillana, Zig-Zag, Piedra de Sol y SM.

A continuación, en la Tabla 1, se presenta la distribución de los textos del corpus en las diferentes asignaturas y cursos bajo estudio.

	Curso			
Asignatura	Sexto	Séptimo	Octavo	Total
LC	93	237	149	479
CN	232	281	286	799
HGCS	299	263	281	843
Total	624	781	716	2121

Tabla 1. Cantidad de textos según nivel y asignatura

Cabe señalar que la distribución total de los textos que se observa en la Tabla 1 corresponde a las instancias que se identificaron en cada asignatura y nivel, eliminando repeticiones. Para evitar la duplicidad de textos en una misma asignatura, pero en diferentes niveles, se realizó un procedimiento que implicó la comparación de todos los textos por asignatura y curso, según la distancia de edición de Levenshtein (1966). Usando un histograma de las distancias, se detectaron pares de textos duplicados o duplicados-cercanos y eliminaron aquellos textos más antiguos según su año de creación cuando pertenecían a un TE del mismo curso o bien se eliminaron aquellos textos de cursos inferiores cuando correspondían a textos presentes en TE de cursos distintos.

En cuanto al tamaño habitual de estos textos, en la Tabla 2 se presenta el promedio de número de palabras de los textos por asignatura y curso.

Curso	LC	HGCS	CN
6º	140.54	256.09	252.23
7º	161.64	275.21	254.79
8º	170.03	357.73	321.43
Total	160.15	295.93	277.90

Tabla 2. Número promedio de palabras de los textos en asignaturas y cursos

7.2 Procesamiento de los textos y cálculo de la LDS

En este trabajo, se utilizó la propuesta de Oya (2011). De este modo, primero, el contenido de cada texto fue entregado en su formato original para la identificación de oraciones. Luego, el conjunto de oraciones resultante

fue usado para extraer todas las palabras o *tokens* en cada oración. Sobre este listado de *tokens* por oración, se realizó un análisis de dependencia sintáctica, el cual generó un árbol de palabras para cada oración del que se extrajo un listado de pares de palabras unidas por la relación Núcleo-Dependiente. Usando estos pares y la posición de cada palabra en el texto, se contabilizó su distancia como la cantidad de palabras existente entre el Núcleo y la Dependiente dentro de la oración. Posteriormente, se calculó el promedio de distancias de cada oración, cuyo valor promediado entre todas las oraciones se reportó como el valor LDS del texto. Estas últimas dos etapas se detallan a continuación en el Procedimiento 1.

```
Input: Conjunto P con pares de oraciones y sus árboles de dependencia sintáctica asociados dentro del texto
Output: Distancia media de dependencia sobre cada palabra y su palabra madre.

ave_LDS = 0
Para cada oración s y su árbol t en P hacer
   dists = 0
   num_pares = 0
   Para cada palabra w en s hacer
      dist_w = separación entre w y palabra madre en t
      incrementar dists en dist_ws
      incrementar num_pares en 1
   fin
   LDS_s = dists/num_pares
   incrementar ave_LDS en LDS_s
fin
reportar ave_LDS/|P| as LDS /*|P| es la cantidad de oraciones del texto t */
```

Procedimiento 1. Cálculo de la LDS para un texto

Inicialmente, el Procedimiento 1 requiere de un listado de oraciones y un árbol de dependencias para cada una de las oraciones, extraído mediante una herramienta computacional de *parsing* (usamos Freeling versión 4.2, ver (Carrera *et al.*, 2008). A partir de este árbol se identifican las palabras relacionadas sobre las cuales se calculan las distancias. Una vez recibida esta información inicial, el algoritmo itera por cada oración del texto identificando cada par de palabras relacionadas y contando la distancia existente entre ambas. Luego de haber procesado todas las oraciones, se obtiene la distancia media entre todos los pares de palabras dentro de las oraciones y finalmente, se calcula el promedio de todas estas distancias medias sobre todas las oraciones del texto.

Para ejemplificar la aplicación del Procedimiento 1 y sin pérdida de generalidad, consideremos un texto compuesto por la siguiente oración:

El_1 $orden_2$ $republicano_3$ es_4 el_5 $resultado_6$ $político_7$ de_8 la_9 $revolución_{10}$ de_{11} $independencia_{12}$

Cuyo árbol de dependencias se muestra en la Figura 1. En la figura se muestran todos los pares de palabras relacionadas dentro del texto. La distancia anotada para cada par se obtiene contando la separación de ambas palabras en el texto. Por ejemplo, dado que la palabra *político* depende sintácticamente de *resultado* y que son contiguas se genera un arco que las conecta y se anota la distancia igual a 1. Para esta oración se contabilizaron 11 arcos (pares de palabras conectadas), cuya suma de distancias es 15, resultando en una LDS de 1,364.

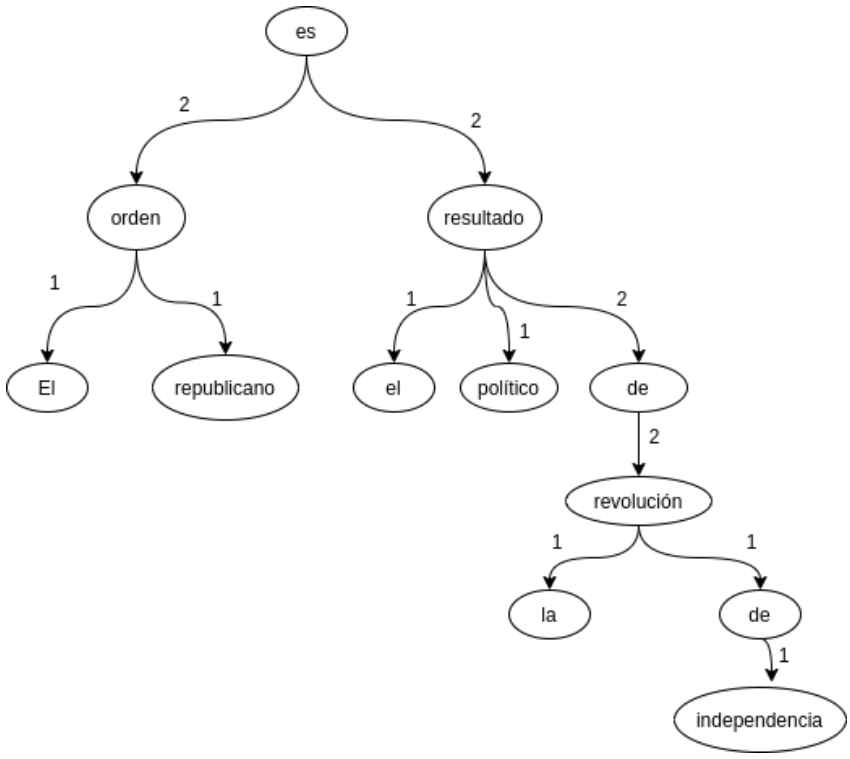


Figura 1. Árbol de dependencias para "El orden republicano es el resultado político de la revolución de independencia".

7.3 Análisis estadístico

Para contrastar cada grupo de textos respecto de los demás grupos se realizó un test de hipótesis no paramétrico. Si bien el mecanismo usado habitualmente para realizar este tipo de comparaciones es el ANOVA (ver por ejemplo Oya, 2011), en este trabajo se utilizó el test de Mann-Whitney, debido a que no se encontró evidencia de normalidad en la distribución de LDS en cada curso tal como se muestra en los resultados del test de Shapiro en el Anexo 2. El test de Mann-Whitney fue aplicado entre cada par de grupos distintos de textos, usando la distribución de todos sus valores de LDS.

8. Resultados

En esta sección se ofrecen los resultados obtenidos por medio de la metodología anteriormente descrita. En primer lugar, se presentan estadísticos descriptivos de LDS registrada por curso y por asignatura. Posteriormente, con el propósito de contrastar los valores de LDS, se presentan los resultados obtenidos con el test de Mann-Whitney en dos niveles, entre cursos y entre asignaturas.

A continuación, en la Tabla 3, se presentan los estadísticos descriptivos de LDS totales, por asignaturas y por cursos.

	LC				HGCS				CN			
Curso	Media	Mediana	DE	CV %	Media	Mediana	DE	CV %	Media	Mediana	DE	CV %
6º	1.993	1.962	0.249	12.49	2.075	2.049	0.252	12.14	2.076	2.052	0.229	11.03
7º	2.188	2.178	0.380	17.37	2.182	2.133	0.314	14.39	2.085	2.052	0.222	10.65
8º	2.075	2.034	0.238	11.47	2.140	2.119	0.189	8.83	2.098	2.061	0.228	10.87
Total	2.115	2.093	0.327	15.46	2.130	2.097	0.259	12.16	2.087	2.055	0.226	10.83

DE= desviación estándar. CV= coeficiente de variación.

Tabla 3. LDS en asignaturas y cursos

Respecto a los valores totales por asignatura, en la Tabla 3 se puede observar que la mayor LDS promedio (2.130) corresponde a la asignatura de HGCS, seguida por LC (2.115) y por último CN (2.087). En cuanto a las medianas, en las tres asignaturas están cercanas a la media. Por otra parte, la mayor dispersión de LDS se encuentra en los textos de LC (15.46 %) y la menor en CN (10.83%). Al considerar cursos y asignaturas, los

promedios mayores corresponden a LC en séptimo (2.188) y HGCS también en séptimo (2.182) y el menor, a sexto de LC (1.993). En relación con las medianas, el valor más alto se encuentra en séptimo de LC (2.178) y el más bajo en sexto de LC (1.962). Por su parte, la mayor dispersión se encuentra en séptimo de LC (17.37 %) y la menor en octavo de HGCS (8.83 %).

Los resultados obtenidos tanto en las medidas de tendencia central como en las de dispersión dan cuenta de una cierta homogeneidad en los datos. No existen grandes diferencias entre los promedios de LDS de los textos analizados y las medianas no se alejan de las medias; de la misma forma, los valores de dispersión dan cuenta de poca variabilidad, en términos generales. Complementando este análisis y guiados por el objetivo de este trabajo, se realizó un análisis estadístico que comparó los promedios entre asignaturas y cursos. Estas comparaciones son presentadas en dos tablas distintas. Con el propósito de describir la proyección según avance curricular, la primera ofrece las comparaciones de los cursos dentro de cada asignatura; con el propósito de describir la eventual variación disciplinar, la segunda muestra las comparaciones por asignatura según curso. Cabe señalar que los resultados presentados corresponden solo a aquellos que son estadísticamente significativos, por tal razón, no todas las comparaciones posibles están incluidas en las tablas.

8.1 Comparación entre cursos de cada asignatura

En la Tabla 4 se presentan los resultados correspondientes a la comparación de los valores de LDS de los cursos por asignatura.

Asignatura	Resultados de comparación entre cursos	Significancia
LC	6º < 7º	0.0000001
	6º < 8º	0.0036020
	8º < 7º	0.0001584
HGCS	6º < 7º	0.0000077
	6º < 8º	0.0000028

Tabla 4. Diferencias estadísticas de LDS promedio entre los cursos en cada asignatura

Como ya se señaló, la Tabla 4 solo ofrece las diferencias que resultaron estadísticamente significativas, lo que explica la ausencia de CN, asignatura en la que no se observaron diferencias en la LDS promedio de los tres cursos analizados (6º, 7º y 8º. En cuanto a LC, se observa que los textos de 6º presentaron una LDS promedio estadísticamente menor que la de los textos de 7º y 8º y, del mismo modo, los textos de 8º presentaron una LDS promedio estadísticamente menor que la de los textos de 7º. Lo anterior quiere decir que la complejidad sintáctica en la asignatura de LC aumenta progresivamente en orden 6to < 8º < 7º. De forma similar, se observa que en la asignatura HGCS, los textos de 6to presentan una LDS promedio estadísticamente menor que los de 7º y 8º y que entre 7º y 8º no se encontraron diferencias estadísticamente significativas, lo que arrojaría una progresión en el orden 6º < 7º = 8º.

8.2 Comparación por curso entre las asignaturas

En la Tabla 5 se presentan los resultados de la comparación de las LDS promedio por cada curso entre las asignaturas. Nuevamente, solo se incluyen los valores cuyas diferencias resultaron estadísticamente significativas.

Curso	Resultados de comparación entre asignaturas	Significancia
6º	LC < HGCS	0.0019160
	LC < CN	0.0009501
7º	CN < LC	0.0000505
	CN < HGCS	0.0001520
8º	LC < HGCS	0.0001217
	CN < HGCS	0.0003101

Tabla 5. Diferencias estadísticas de LDS promedio entre asignaturas por curso

Como se observa en la Tabla 5, en 6º, los textos de la asignatura LC presentan una LDS promedio estadísticamente menor que la de los textos de HGCS y CN y entre las asignaturas HGCS y CN no se observaron diferencias estadísticas. En cuanto a 7mo, los textos de CN presentan una LDS promedio menor que la de los textos de LC y HGCS. Entre las asignaturas LC e HGCS, por su parte, no se encuentran diferencias estadísticamente significativas. Respecto a 8º, se observa que los textos de las asignaturas LC y CN presentan promedios de LDS estadísticamente menores que los de la asignatura HGCS, mientras que entre las asignaturas LC y CN no se encontraron diferencias estadísticamente significativas.

9. Discusión

Los resultados obtenidos a partir del análisis automático nos han permitido obtener datos empíricos relevantes respecto de las características sintácticas de los TE chilenos de educación básica. Más específicamente, ha sido posible comparar los niveles de CS entre cursos y asignaturas por medio de la LDS, propuesta que entendemos como una forma eficiente de operacionalizar el constructo. En este mismo sentido y a diferencia de lo que han planteado algunos investigadores (Hawkins, 1990; Arnold *et al.,* 2000), el análisis automático realizado ofrece pruebas de que la CS no puede ser reducida solo a la cantidad de elementos de una oración; también es necesario considerar la relación entre tales elementos para poder aproximarse al nivel de complejidad de una construcción sintáctica determinada (Frantz *et al.,* 2015).

Como ya se señaló, la hipótesis inicial señalaba que la CS variaría entre las asignaturas y se incrementaría a medida que aumenta el nivel escolar. Sin embargo, los resultados generales de nuestro estudio no están completamente en línea con nuestras expectativas. En primer lugar, llama la atención los bajos promedios de LDS, los cuales no exceden las 2.2 palabras en ninguna asignatura. Este bajo nivel de LDS se evidencia en el ejemplo 3, extracto de un texto de LC de sexto año básico:

(3) Los mitos griegos eran relatos orales originarios de tiempos remotos. Fueron recogidos por Homero, en la *Ilíada* y la *Odisea* y por Hesíodo, en diversas historias, alrededor del año setecientos antes de Cristo. Posteriormente, los romanos adoptaron los mitos más importantes de la mitología griega y nombraron a los dioses de manera diferente. El escritor latino Ovidio narró en su libro *Metamorfosis* más de doscientos mitos.

En el ejemplo se puede apreciar el uso preferente de relaciones de baja LDS. En un extracto compuesto por cuatro oraciones, casi todas de considerable extensión, podemos observar la ausencia de cláusulas subordinadas como relativas o adverbiales, las cuales, evidentemente, aumentan la LDS. Los complementos circunstanciales, como los de tiempo, se estructuran en función de frases nominales que se ubican al final de la oración, lo cual no afecta el carácter adyacente del núcleo oracional y sus argumentos (*alrededor del año setecientos antes de Cristo*). Del mismo modo, el complemento circunstancial temporal se traduce en el uso de un adverbio de modo (*posteriormente*) que, tratándose de una sola palabra, colocada en la periferia izquierda de la oración, no afecta la LDS entre núcleos y dependientes, todo lo cual puede explicar los bajos promedios de LDS hallados en nuestra investigación. Estos bajos promedios de LDS y la homogeneidad de los datos, considerando una DE general que no supera el 0.4, nos permite afirmar que la CS de los textos de nuestro corpus, en general, es bastante baja.

Respecto de las implicancias de los resultados generales, como ya se señaló en el apartado teórico, diversos estudios han demostrado que el efecto de la LDS se comienza a manifestar en el procesamiento y en la comprensión, a partir de una distancia de tres palabras (Grodner & Gibson 2005; Bartek *et al.,* 2011; Nicemboin *et al.,* 2015, Kleijn, 2018). Más precisamente, las construcciones sintácticas con LDS mayor a cinco serían más difíciles de procesar y de comprender debido al efecto que tal distancia tiene en la memoria de trabajo. En virtud de lo anterior, la LDS promedio observada en términos generales en el corpus analizado nos lleva a inferir que los TE a los cuales se enfrentan las y los estudiantes chilenos de educación básica no representan una dificultad mayor en términos sintácticos. Estos datos resultan relevantes dado que la CS, operacionalizada en términos de LDS, incide directamente en el procesamiento, la comprensión y el aprendizaje a partir del texto (Gibson, 1998, 2000; Grodner & Gibson, 2005; Futrell *et al.,* 2019).

En cuanto a la CS de los textos en los distintos cursos de cada asignatura, se esperaba que los promedios de LDS aumentaran a medida que avanzara el nivel de escolaridad. Tales expectativas se basan en la revisión de los términos de referencia establecidos por el MINEDUC Chile (2018). Con respecto a esto, los resultados obtenidos indican que las asignaturas HGCS y LC repiten un patrón en cuanto al comportamiento de los promedios. En ambas sucede que el promedio de LDS de los textos no aumenta en el orden que aumenta el curso, dado que el menor promedio es en 6º, luego 8º y el más alto corresponde a 7º. Llama la atención también el caso de CN, pues no se encontraron diferencias en los promedios de los diferentes niveles de escolaridad.

Estos resultados indican que no se está cumpliendo con la progresión en la CS de los textos que se explicita en los requerimientos del Estado Chileno. Esto implica, por un lado, que estos textos podrían no constituir

el material adecuado para mejorar la competencia lectora, pues no se complejizan al tiempo que las y los estudiantes avanzan en su escolaridad y desarrollo cognitivo, y, por otro lado, no contribuirían a estimular a los lectores, pues los textos no desafían las capacidades intelectuales de los aprendientes, al menos, en términos de CS. Los bajos promedios de CS hallados implicarían, además, que el estudiantado no está expuesto, en sus materiales de estudio, a estructuras sintácticas complejas, lo que podría tener consecuencias no solo en la lectura sino también en el desarrollo de la producción escrita.

Una segunda mirada comparativa a los datos obedece al hecho de que las disciplinas comunican el conocimiento de maneras distintas (Hyland, 2000; Wheelahan, 2010), por lo que se esperaba no solo que existieran diferentes LDS promedio por cada curso entre las diferentes asignaturas, sino que también -y con base en los resultados obtenidos por Graesser *et al.* (2011) en un estudio similar- un orden jerárquico, en el que HGCS tendría la mayor LDS promedio y CN, la menor. A pesar de lo anterior, los resultados obtenidos no permiten observar ningún patrón de distribución claro, puesto que en cada curso las asignaturas que presentan mayor y menor LDS promedio son distintas y en algunos casos no existen diferencias entre las que presentan la mayor y en otros, entre las que presentan la menor. Sin embargo, una mirada más pormenorizada de la situación permite distinguir una tendencia de LC y de CN a una menor LDS promedio y una tendencia de HGCS a una mayor. Estas tendencias pueden ser relacionadas con las diferencias encontradas por Graesser *et al*. (2011). A partir de estos resultados y si bien los promedios generales de LDS son homogéneamente bajos, sería posible sostener que los textos de la asignatura de HGCS representan mayor dificultad a las y los estudiantes en términos de CS.

Por otra parte, y a pesar de no haber encontrado un patrón que permita distinguir con claridad las diferencias discursivas determinadas por la disciplina, estos resultados también representan evidencia para la hipótesis de la variación disciplinar, la que algunos autores han identificado como una variación en el registro académico, determinada por las variables contextuales (Frantz *et al.,* 2015; Schleppegrell, 2004) y otros han asociado a las diferentes formas en que el conocimiento es recontextualizado (Christie, 2002; Ibáñez *et al.,* 2017; 2018).

10. Conclusiones

El objetivo que orientó esta investigación fue comparar, por medio de una herramienta de análisis automático, la CS de los textos utilizados para comunicar el conocimiento en TE de tres asignaturas de educación básica (LC, CN y HGCS). Para ello, se recolectó un corpus compuesto por 2121 instancias del género Exposición de Contenido (Ibáñez *et al.,* 2017) presentes en los TE de sexto, séptimo y octavo básico, entregados por el Estado de Chile a las y los estudiantes de colegios públicos. Tales instancias fueron sometidas a un análisis automático, por medio de un algoritmo que permite determinar la dependencia sintáctica entre los constituyentes de una oración y, del mismo modo, calcular su promedio de LDS. Los resultados revelaron que la LDS promedio de los textos analizados, correspondientes a diferentes cursos y asignaturas es homogéneamente baja. Del mismo modo, se observó que no existe un patrón de complejización a medida que avanzan los cursos. También quedó en evidencia que, si bien no existen patrones disciplinares para determinar que existen asignaturas con mayor CS, sí existe una tendencia que sitúa a HGCS como la más compleja de ellas.

De acuerdo con los resultados de esta investigación, se puede afirmar que la CS de los textos analizados no debiera representar mayor dificultad para las y los estudiantes. Este hallazgo es de gran relevancia, pues si bien no existe controversia respecto de que la dificultad de los textos está determinada por factores lingüísticos de diferentes niveles, así como también por factores extralingüísticos, la CS es frecuentemente considerada un indicador central de lecturabilidad (McNamara *et al.,* 2014). Además de lo anterior, quedó en evidencia que, al menos en términos de CS, los TE entregados por el Estado de Chile no cumplirían con los lineamientos declarados por el propio Ministerio de Educación chileno (MINEDUC Chile, 2015, 2018a, 2018b) en cuanto a su incremento gradual de complejidad.

Otro aspecto a relevar de este estudio es que tradicionalmente, el análisis de la CS se ha realizado de forma manual para el español, por lo que no existen precedentes de su cálculo en forma automática. El enfoque de operacionalización de CS en términos de LDS así lo permite y representa una manera eficiente de realizar la tarea y de gran utilidad para determinar la dificultad que en términos sintácticos, los textos podrían representar para los lectores. Sin lugar a dudas, este tipo de estudios tiene una proyección natural en la corroboración psicolingüística, por lo que un siguiente paso lo constituye comprobar, mediante pruebas de procesamiento y comprensión de textos escritos el efecto real de los promedios de LDS identificados en el desempeño de las y los estudiantes.

Agradecimientos

Este trabajo se realizó bajo el financiamiento de la Agencia Nacional de Investigación y Desarrollo de Chile por medio de los proyectos Fondecyt 1201440 y Fondecyt 11200826.

Bibliografía

Altamirano, P., Godoy, G., Manghi, D. & Soto, G. (2014). Analizando los textos de Historia, Geografía y Ciencias Sociales: La configuración multimodal de los pueblos originarios. *Estudios Pedagógicos, 40*(1), 263-280. https://doi.org/10.4067/s0718-07052014000100016

Aravena, S. & Hugo, E. (2016). Desarrollo de la complejidad sintáctica en textos narrativos y explicativos escritos por estudiantes secundarios. *Lenguas Modernas* (47), 9-40. https://revistas.uchile.cl/index.php/LM/article/view/45181

Arnold, J., Wasow, T., Losongco, A. & Ginstrom, R. (2000). Heaviness vs. Newness: The Effects of Structural Complexity and Discourse Status on Constituent Ordering. *Language, 17*(1), 28-55. https://doi.org/10.1353/lan.2000.0045

Bailey, A., Butler, F., Stevens, R. & Lord, C. (2007). Further specifying the language demands of school. En A. Bailey (Ed.), *The language demands of school: Putting academic English to the test* (pp. 103-156). Yale University Press.

Ballesteros, M., Bohnet, B., Mille, S., & Wanner, L. (2016). Data-driven deep-syntactic dependency parsing. *Natural Language Engineering, 22*(6), 939-974. https://doi.org/10.1017/S1351324915000285

Bañados, E. (2007). Integrando las tecnologías de información y comunicación en el currículum, como recurso pedagógico complementario al texto escolar en la enseñanza-aprendizaje de idiomas extranjeros. En *Primer Seminario Internacional de textos escolares SITE 2006*. Disponible en https://bibliotecadigital.mineduc.cl

Bartek, B., Lewis, R., Vasishth, S. & Smith, M. (2011). In Search of On-Line Locality Effects in Sentence Comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition, 37*(5), 1178-1198. https://doi.org/10.1037/a0024194

Bartolomé, R. 2021. Estudio comparativo de los índices de madurez sintáctica entre las generaciones pre y post internet. *Círculo de Lingüística Aplicada a la Comunicación*, (88), 83-106. https://dx.doi.org/10.5209/clac.78299

Carrera, J., Castellón, I., lloberes, M., Padró, L. & Tinkova, N. (2008). Dependency Grammars in FreeLing. *Procesamiento de LC Natural*, 41, 21-28.

Chatzipanagiotidis, S., Giagkou, M., & Meurers, D. (2021). Broad linguistic complexity analysis for Greek readability classification. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 48-58).

Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 740-750). https://doi.org/10.3115/v1/D14-1082

Chomsky, N. (1970). *Aspectos de la teoría de la sintaxis*. Aguilar.

Chomsky, N. (1978). *Estructuras sintácticas*. Siglo XXI.

Choppin, A. (2000). Pasado y presente de los manuales escolares. En J. Berrio (Ed.), *La cultura escolar de Europa. Tendencias históricas emergentes* (pp.107-165). Biblioteca Nueva.

Christie, F. (1998). Science and apprenticeship. The pedagogic discourse. En J. Martin & R. Veel (Eds.), *Reading science. Critical and functional perspectives on discourse of science* (pp. 152-180). Routledge.

Christie, F. (2002). *Classroom discourse analysis: A functional perspective*. Continuum.

Crespo, N. Alvarado, C. & Meneses, A. (2013). Desarrollo sintáctico: Una medición a partir de la diversidad clausular. *Logos. Revista de Lingüística, Filosofía y Literatura, 23*(1), 80-101. https://revistas.userena.cl/index.php/logos/article/view/197

Crespo, Alfaro & Góngora (2011). La medición de la sintaxis. Evolución de un concepto. *Onomázein, 24*(2), 155-172. http://revistaaisthesis.uc.cl/index.php/onom/article/view/30967

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition, 109*(2), 193-210. https://doi.org/10.1016/j.cognition.2008.07.008

Dyer, C., Ballesteros, M., Ling, W., Matthews, A., & Smith, N. A. (2015). Transition-Based Dependency Parsing with Stack Long Short-Term Memory. En *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 334-343). https://doi.org/10.3115/v1/P15-1033

Falkenjack, J., Mühlenbock, K. H., & Jönsson, A. (2013). Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)* (pp. 27-40).

Falkenjack, J., Santini, M., & Jönsson, A. (2016). An exploratory study on genre classification using readability features. En *Proceedings of the Sixth Swedish Language Technology Conference (SLTC 2016),* Umeå, Sweden. https://doi.org/10.13140/RG.2.2.33356.21120

Fedorenko, E., Woodbury, R. & Gibson, E. (2013). Direct Evidence of Memory Retrieval as a Source of Difficulty in Non-local Dependencies in Language. *Cognitive Science, 37*, 378-394. https://doi.org/10.1111/cogs.12021

Ferreira, F. (1991). Effects of Length and Syntactic Complexity on Initiation Times for Prepared Utterances. *Journal of Memory and Language, 30*(2), 2110-2233. https://doi.org/10.1016/0749-596X(91)90004-4

Frantz, R., Starr, L. & Bailey, A. (2015). Syntactic Complexity as an Aspect of Text Complexity. *Educational Researcher, 44*(7), 387-393. https://doi.org/10.3102/0013189X15603980

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences, 112*(33), 10336-10341. https://doi.org/10.1073/pnas.1502134112

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers*), pp 32–42, Minneapolis, Minnesota. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/N19-1004

Futrell, R., Gibson, E. & Levy, R. (2020). Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Cognitive Science, 44*, 2-54. https://doi.org/10.1111/cogs.12814

Gernsbacher M. A. (1989). Mechanisms that improve referential access. *Cognition, 32*(2), 99–156. https://doi.org/10.1016/0010-0277(89)90001-2

Gibson, E. (1998). Linguistics complexity: Locality of syntactic dependencies. *Cognition, 68*(1), 1-76. https://doi.org/10.1016/S0010-0277(98)00034-1

Gibson, E. (2000). The Dependency Locality Theory: A Distance-Based Theory of Linguistics Complexity. En A. Marantz, Y. Miyashita, & W. O´Neil, (Eds.), *Image, language, brain: Papers for the first mind articulation project symposium* (pp. 95-126). MIT Press.

Gili Gaya, S. (1972). El pretérito de negación implícita. En *Studia Hispanica in honorem R*. Gredos, tomo I, 251-6.

Givón, T. (1991). Markedness in grammar: distributional, communicative and cognitive correlates of syntactic structures. *Studies in Language, 15*(2), 335-370. http://dx.doi.org/10.1075/sl.15.2.05giv

Givón, T. (2009). *The Genesis of Syntactic Complexity*. John Benjamins Publishing Company. http://dx.doi.org/10.1075/z.146

Goldman, S. (1997). Learning from texts: Reflections on the past and suggestions for the future. *Discourse Processes, 23*, 357-398. http://dx.doi.org/10.1080/01638539709544997

Graesser, A.C., McNamara, D.S., & Kulikowich, J.M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223–234. https://doi.org/10.3102/0013189X11413260

Graesser, A.C. & McNamara D.S. (2011). Computational analyses of multilevel discourse comprehension. Topics in *Cognitive Science, 3*, 371-98. https://doi.org/10.1111/j.1756-8765.2010.01081.x

Grodner, D. & Gibson, E. (2005). Consequences of the Serial Nature of Linguistics Input for Sentencial Complexity. *Cognitive Science, 29*, 261-290. https://doi.org/10.1207/s15516709cog0000_7

Gysling, J. & Meckes, L. (2011). "Estándares de aprendizaje en Chile: mapas de progreso y logro SIMCE 2002 a 2010", PREAL Serie Documentos Nº 54. Inter-American Dialogue.

Hall, D., Durrett, G., & Klein, D. (2014). Less grammar, more features. En *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 228-237). https://doi.org/10.3115/v1/P14-1022

Hawkins, J. (1990). A Parsing Theory of Word Order Universals. *Linguistics Inquiry, 21*(2), 223-261. http://www.jstor.org/stable/4178670

Hawkins, J. (1994). *A Performance Theory of Order and Constituency*. Cambridge University Press. https://doi.org/10.1017/CBO9780511554285

Herrera Lima, M.E. (1991). Madurez sintáctica en escolares de Ciudad de México. Análisis preliminar. En H. López Morales (Ed.), *La enseñanza del español como lengua materna* (pp. 155-169). Universidad de Puerto Rico.

Hunt, K. W. (1965). Grammatical Structures Written at Three Grade Levels. National Council of Teachers of English Research Report N° 3. National Council of Teachers on English, Urbana.

Hunt, K.W. (1970). Syntactic Maturity in Schoolchildren and Adults. *Monographs of the Society for Research in Child Development, 35*(1), 1-67.

Hyland, K. (2000). *Disciplinary Discourses: Social Interactions in Academic Writing*. Longman.

Hyland, K. (2004). *Disciplinary Discourses: Social Interactions in Academic Writing*. University of Michigan Press. https://doi.org/10.3998/mpub.6719

Ibáñez, R., Moncada, F., Cornejo, F., & Arriaza, V. (2017). Los Géneros del Conocimiento en Textos Escolares de Educación Primaria. *Calidoscopio, 15*(1), 462-476. http://revistas.unisinos.br/index.php/calidoscopio/article/view/cld.2017.153.06

Ibáñez, R., Moncada, F., & Arriaza, V. (2018). Recontextualización del conocimiento en textos escolares chilenos. *Revista Signos. Estudios de Lingüística, 51*(98), 430-456. http://dx.doi.org/10.4067/S0718-09342018000300430

Ibáñez, R., Moncada, F., & Cárcamo, B. (2019). Coherence Relations in Primary School Texts Books: Variation Across School Subjects. *Discourse Processes, 56*, 764-785. https://doi.org/10.1080/0163853X.2019.1565278

Kasule, D. (2011). Textbook Readability and ESL Learner. *Reading and Writing, 2*, 63-76. http://dx.doi.org/10.4102/rw.v2i1.13

King, J. & Just, M. (1991). Individual Differences in Syntactic Processing: The Role of Working Memory. *Journal of Memory and Language, 30*, 580-602. https://doi.org/10.1016/0749-596X(91)90027-H

Kitaev, N., & Klein, D. (2018). Constituency Parsing with a Self-Attentive Encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2676-2686). http://dx.doi.org/10.18653/v1/P18-1249

Kleijn, S. (2018). *Clozing in on readability: how linguistics features affect and predict text comprehension and on-line processing*. LOT Publications.

Koo, T., Carreras, X., & Collins, M. (2008). Simple semi-supervised dependency parsing. In *Proceedings of ACL-08: HLT* (pp. 595-603).

Kübler, S., McDonald, R., & Nivre, J. (2009). Dependency parsing. *Synthesis lectures on human language technologies, 1*(1), 1-127. https://doi.org/10.2200/S00169ED1V01Y200901HLT002

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady, 10*(8), 707-710.

Levy, R. & Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language, 68*(2), 199-222. https://doi.org/10.1016/j.jml.2012.02.005

Lewis, R. & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science, 29*, 375-419. https://doi.org/10.1207/s15516709cog0000_25

Li, H., Zhang, Z., Ju, Y., & Zhao, H. (2018). Neural character-level dependency parsing for Chinese. En *Thirty-Second AAAI Conference on Artificial Intelligence*. Disponible en https://ojs.aaai.org/index.php/AAAI/article/view/12002

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science, 9*(2), 159-191. http://dx.doi.org/10.17791/jcs.2008.9.2.159

Marcheggiani, D., Frolov, A., & Titov, I. (2017). A Simple and Accurate Syntax-Agnostic Neural Model for Dependency-based Semantic Role Labeling. En *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)* (pp. 411-420). http://dx.doi.org/10.18653/v1/K17-1041

Martin, J. & Rose, D. (2008). *Genre Relations: Mapping Culture*. Equinox.

Martin, J. & Rose, D. (2013). Pedagogic Discourse: Contexts of Schooling. *RASK: International Journal of Language and Communication, 38*, 219-264.

McNamara, D. S. & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes, 22*(3), 247–288. https://doi.org/10.1080/01638539609544975

McNamara, D. S. (2004). Aprender del texto: Efectos de la estructura textual y las estrategias del lector. *Revista Signos. Estudios de Lingüística, 37*(55), 19-30. http://dx.doi.org/10.4067/S0718-09342004005500002

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014) Automated Evaluation of Text and Discourse with Coh-Metrix. Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511894664

MINEDUC (2015). Bases Curriculares. 7mo básico a 2do medio. Disponible en https://www.curriculumnacional.cl/portal/Documentos-Curriculares/.

MINEDUC (2018a). Bases Curriculares. Primero a Sexto Básico. Disponible en https://www.curriculumnacional.cl/portal/Documentos-Curriculares/.

MINEDUC (2018b). ¿Qué debemos saber sobre los textos escolares? Disponible en https://www.supereduc.cl/contenidos-de-interes/que-debemos-saber-sobre-los textosescolares

Möenne, G. & López, L. (2007). Oportunidades que ofrecen las TICs como apoyo a los textos escolares. En *Primer Seminario Internacional de textos escolares SITE 2006*. Disponible en https://bibliotecadigital.mineduc.cl

Nicemboin, B., Vasishth, S., Gattei, C., Sigman, M. & Kliegl, R. (2015). Working memory differences in long-distance dependency resolution. *Frontiers in Psychology, 6*, 2-16. https://doi.org/10.3389/fpsyg.2015.00312

Nir, B. & Berman, R. (2010). Complex syntax as a window on contrastive rhetoric. *Journal of Pragmatics, 42*(3), 744-765. https://doi.org/10.1016/j.pragma.2009.07.006

Nivre, J., & McDonald, R. (2008). Integrating graph-based and transition-based dependency parsers. In *Proceedings of ACL-08: HLT* (pp. 950-958).

Olivares, P. (2007). Concepto de nación e identidad nacional: Una approache a través de las políticas educativas y de la enseñanza de la Historia de Chile (Siglos XIX-XX). En MINEDUC (Ed.), *Acta del Primer Seminario Internacional de Textos Escolares* (pp.161-165). Mineduc-UNESCO.

Oteiza, T. (2009). Cómo es presentada la historia contemporánea en los libros de textos chilenos para la escuela media. *Discurso & Sociedad, 3*(1), 150-174.

Oya, M. (2011). Syntactic dependency distance as sentence complexity measure. En *Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics* (pp. 313-316).

Peñaloza, C., Araya, C. & Coloma, C.J. (2017). Desarrollo de la complejidad sintáctica en recontados narrativos de niños preescolares y escolares. *Logos. Revista de Lingüística, Filosofía y Literatura, 27*(2), 333-348. https://doi.org/10.15443/RL2726.

Pilán, I., Volodina, E., & Johansson, R. (2014). Rule-based and machine learning approaches for second language sentence-level readability. En *Proceedings of the ninth workshop on innovative use of NLP for building educational applications* (pp. 174-184). http://dx.doi.org/10.3115/v1/W14-1821

Poulsen, M. & Gravgaard, A. (2016). Who did what to whom? The relationship between syntactic aspects of sentence comprehension and text comprehension. *Scientific Studies of Reading, 20*(4), 325-338. http://dx.doi.org/10.1080/10888438.2016.1180695

Ramírez, T. (2002). El Texto Escolar como Objeto de Reflexión e Investigación. *Docencia Universitaria, 3*(1), 101-124.

Rickford, J., Denton, M., Wasow, T. & Espinoza, J. (1995). Syntactic Variation and Change in Progress: Loss of the Verbal Coda in Topic-Restricting As Far As Constructions. *Language, 71*(1), 102-131. https://doi.org/10.2307/415964

Rodríguez Fonseca, L. (1991). Índices de madurez sintáctica en escolares puertorriqueños de escuela primaria. En H. López Morales (Ed.), *La enseñanza del español como lengua materna* (pp. 133-143). Universidad de Puerto Rico.

Rojas, D., Ibáñez, R., Moncada, F., & Santana, A. (2020). Los Géneros del Conocimiento en el texto escolar de Lenguaje y Comunicación: Un análisis semiautomático de su lecturabilidad. *RLA. Revista de Lingüística Teórica y Aplicada, 58*(2), 41-67. https://doi.org/10.29393/RLA58-14GCDR40014

Rojas, D. (2021). *Efecto de la complejidad sintáctica en la comprensión de estudiantes de Octavo Año Básico: una aproximación a la lecturabilidad del Texto Escolar*. [Tesis de Magíster, Pontificia Universidad Católica de Valparaíso]. http://repositorio.conicyt.cl/bitstream/handle/10533/249935/Tesis_Rojas_Villarroel.pdf?sequence=1

Rose, D. (2014). Analyzing pedagogic discourse: An Approach from genre and register. *Functional Linguistics, 1*, 11. https://doi.org/10.1186/s40554-014-0011-4

Schleppegrell, M. J. (2004). *The Language of Schooling*. Lawrence Erlbaum.

Sedano, M. (2011). *Manual de gramática del español, con especial referencia al español de Venezuela*. Consejo de Desarrollo Científico y Humanístico, Universidad Central de Venezuela.

Sánchez, V. & De Mier, V. (2017). Syntactic Complexity in Narratives Written by Spanish Heritage Speakers. *Vigo International Journal of Applied Linguistics*, (14), 125-148.

Santana, A. Ibáñez, R. Moncada, F. & Zamora, J. (2021). Causal Connective Expressions in Textbooks written in Spanish: a comparative study of four primary school subjects. *Journal of Pragmatics, 182*, 104-117. https://doi.org/10.1016/j.pragma.2021.06.010

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition, 116*, 71-86. https://doi.org/10.1016/j.cognition.2010.04.002

Tesnière, L. (2015). *Elements of structural syntax.* John Benjamins Publishing Company. https://doi.org/10.1075/z.185

Vasishth, S. & Drenhaus, H. (2011). Locality in German. *Dialogue Discourse, 2*, 59-82.

Vásquez, I. (1991). Índices de madurez sintáctica en estudiantes puertorriqueños de escuela superior. En H. López Morales (Ed.), *La enseñanza del español como lengua materna* (pp. 145-153). Universidad de Puerto Rico.

Véliz, M. 1988. Evaluación de la madurez sintáctica en el discurso escrito. *Revista de Lingüística Teórica y Aplicada, 26*, 105-141.

Véliz, M. 1999. Complejidad Sintáctica y modo del discurso. *Revista de Estudios Filológicos, 34,* 181-192. http://dx.doi.org/10.4067/S0071-17131999003400013

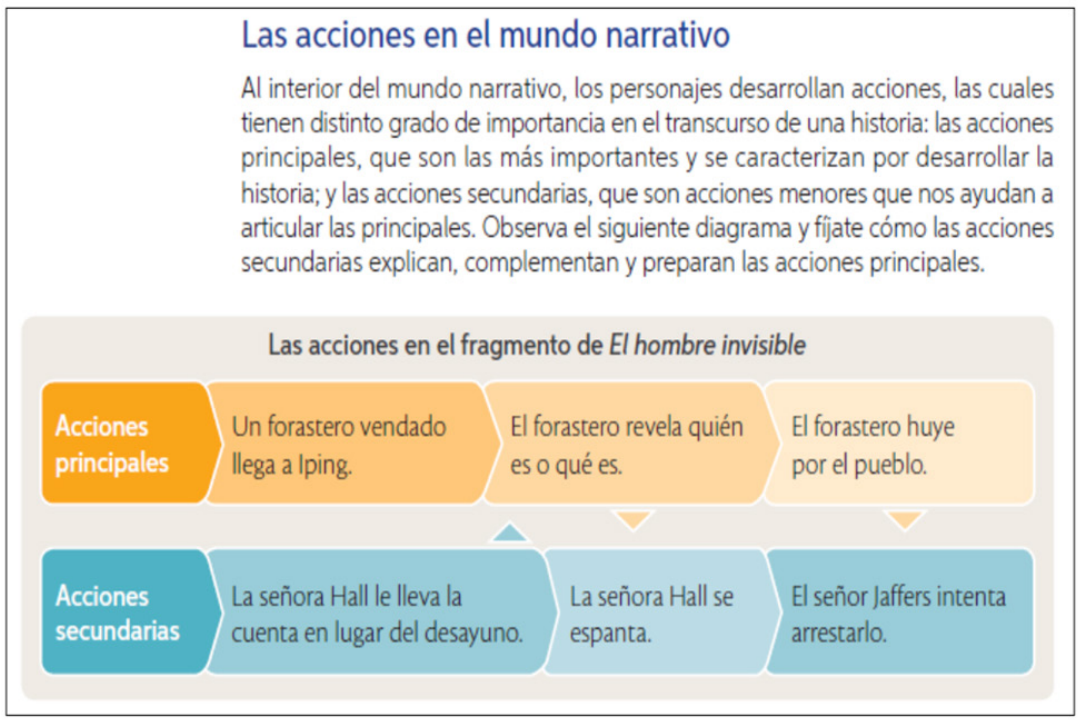
Watanabe, T., & Sumita, E. (2015). Transition-based neural constituent parsing. En *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1169-1179). http://dx.doi.org/10.3115/v1/P15-1113

Wheelahan, L. (2010). *Why knowledge matters in curriculum.* Roudledge. https://doi.org/10.4324/9780203860236

Yamada, H., & Matsumoto, Y. (2003). Statistical dependency analysis with support vector machines. In *Proceedings of the eighth international conference on parsing technologies* (pp. 195-206).

Anexos

Anexo I: Género Exposición de Contenidos



Anexo II. Test de Shapiro

Asignatura	Curso	Shapiro Test	Hipótesis de Normalidad
CN	6	9.877e-12	Rechazada
	7	2.648e-20	Rechazada
	8	1.177e-18	Rechazada
HGCS	6	6.403e-37	Rechazada
	7	4.284e-33	Rechazada
	8	5.188e-37	Rechazada
LC	6	2.435e-34	Rechazada
	7	2.747e-33	Rechazada
	8	2.017e-33	Rechazada

Tabla 6. Test de normalidad de Shapiro para LDS en cada asignatura junto con el resultado de la prueba de hipótesis.