Pontificia Universidad Católica de Chile
Facultad de Letras

Léxico disponible de estudiantes universitarios de Educación Básica y Letras Hispánicas: un estudio sociolingüístico comparativo y análisis de técnicas de recolección de datos

Tesis presentada como requisito parcial para obtener el grado de Doctor en Lingüística

José Alejandro Martínez-Lara

Directora: Dra. María Natalia Castillo Fadić
Codirectora: Dra. Inmaculada Clotilde Santos Díaz

Santiago de Chile
2 de octubre de 2023

Resumen

Esta tesis doctoral examina el léxico disponible (en adelante LD) –es decir, el conjunto de palabras que viene más rápido a la mente cuando se activa un tema específico a través de un estímulo verbal (Michéa, 1953; Gougenheim *et al.*, 1964; López Morales, 1995-1996; Gómez Molina, 2021, entre otros) – referido a los centros de interés (CI) 01. *La lectura*, 02. *El profesor*, 03. *La educación*, 04. *Juegos y distracciones*, 05. *La escuela*, 06. *Habilidades docentes*, 07. *Partes del cuerpo* y 08. *Comidas y bebidas*, de universitarios de Educación Básica (EB) y Letras Hispánicas (LH) de la Pontificia Universidad Católica de Chile. Con base en este tema, el objetivo de esta investigación es conocer cuánto y cuál es el LD, recolectado tanto en formato papel como digital, con la finalidad de determinar, a través de análisis comparativos, las convergencias y divergencias de los lexicones de los estudiantes de EB y LH, así como evaluar la técnica alternativa de recogida de datos a través de un instrumento electrónico. Para esto, se examinó, cuantitativa y cualitativamente, un corpus producido por 264 participantes, organizado en tres muestras: 1) Educación Básica; 2) Letras Hispánicas, estas dos fueron recogidas a través de instrumentos en papel; y 3) Letras Hispánicas, colectada mediante la página web *ad hoc*. Los resultados cuantitativos generales indican que los CI más productivos, ricos y compactos son –ordenados según sus respectivos rangos– el 07, 08 y 05. En tanto que los análisis inferenciales señalan que son significativas las relaciones entre *Sexo* y el CI05, así como *Año de curso* y CI08, a favor de las variantes mujeres y 1.er Año, respectivamente. En cuanto a la variable *Carrera*, específicamente, debe destacarse que las medias globales de LH (19,71) y EB (16,51) se diferencian por 3,2 puntos, y muestra una significancia de p = ,000 en el CI01, siendo las palabras más disponibles: *libro*, *letra*, *leer*, *autor* y *palabra*. Por su parte, la variable *Formato de pruebas* exhibe promedios de palabras globales, entre la muestra en papel (19,71) y la digital (19,23), bastante similares, con una distinción de apenas 0,48 lexías. Asimismo, se aprecia una significación del factor en el CI06 (p = 0,32); los vocablos más disponibles de este grupo son: *empatía*, *comprensión*, *vocación*, *paciencia* y *enseñar*. Por último, al analizar los grafos del CI01, se determinó que los tres grupos presentan asociaciones semánticas parecidas, en las que se evidencian los recursos de meronimia, afinidad semántica y proximidad visual. En conclusión, los análisis, cuantitativos y cualitativos, han permitido demostrar que existen diferencias léxico-métricas significativas entre las comunidades discursivas (Parodi, 2004; Bolívar, 2013). Asimismo, se ha logrado evaluar un instrumento digital que puede recoger datos léxicos de manera acertada, por lo que debe ahondarse más sobre este tópico.

Palabras clave: disponibilidad léxica, léxico disponible, léxico de universitarios, sociolingüística, análisis de grafos.

TESI[ANDO]

No hay tres que no tenga cuatro…

¿La tercera no era la última?

No hay mal que por bien no venga

O

No hay cuatro que de tres no venga…

Entonces, la cuarta bienvenida sea…

La cuarta será la vencida.

¿La cuarta será la última?

Al menos, la pude terminar.

Al Dios, Trino y Uno, en quien me sostengo.

A Ennio, Carmen Natalia y Esquia Tocuyo.

Índice general

Índice de figuras

Índice de tablas

Índice de gráficos

Introducción

> "Las palabras siempre han desempeñado un papel importante en la vida humana y han sido y son el epicentro que atraviesa una serie de "hechos humanos"" (Otaola, 2004: 2)

Esta tesis doctoral aborda el léxico disponible –entendido como el conjunto de palabras que viene más rápido a la mente de los interactuantes cuando se activa un tema específico (Michéa, 1953; Gougenheim *et al.*, 1964; López Morales, 1995-1996; entre otros) – de estudiantes de los programas de pregrado de Educación Básica y Letras Hispánicas de la Pontificia Universidad Católica de Chile, desde una perspectiva sociolingüística, con énfasis en los factores *Carrera* y *Formato de pruebas*.

Los léxicos disponibles ofrecen información valiosa sobre la actualización del vocabulario de los diferentes grupos sociales, por lo que su elaboración constituye un pilar necesario en la orientación de decisiones sobre aspectos de índole lingüística. En este sentido, los estudios léxico-métricos cobran especial relevancia en el ámbito educativo, porque gracias a ellos puede evaluarse el componente lingüístico del alumnado. Además, con base en ellos, puede seleccionarse empíricamente el conjunto de palabras que deberían contener los materiales destinados a alcanzar los diferentes objetivos pedagógicos, entre otras razones (cf. Castillo Fadić, 2021a). Al respecto, ya Valencia y Echeverría (1999) daban cuenta de la necesidad de indagar, cuantitativa y cualitativamente, acerca de la conformación y naturaleza del léxico de los hablantes a fin de detectar sus posibles deficiencias y, a partir de ello, elaborar estrategias didácticas para suplirlas. Esta aseveración se fundamenta sobre el supuesto de que los conocimientos –ya sean de tipo científico o disciplinar, representacional o construido (Osses y Jaramillo, 2008) – se adquieren mediante el lenguaje y, en consecuencia, se reflejan a través de él.

En esta línea argumental, el léxico disponible (en adelante LD) se presenta como un instrumento eficaz en la evaluación y descripción de la realidad lingüística de una comunidad, puesto que se trata de un vocabulario potencial, actualizable en la medida en que lo requiera el contexto comunicativo. Asimismo, muestra no solo la riqueza de palabras, sino también la prolijidad de la ortografía, las redes asociativa, la adquisición de conocimientos, entre otros aspectos. En palabras de Gómez Molina (2021: 208), el LD "es el conjunto de unidades léxicas de contenido semántico concreto que el hablante de una lengua puede utilizar de forma inmediata", para lo cual el tópico al que pertenece una pieza léxica X debe estar activado durante la enunciación. Este tipo de vocabulario es analizado desde el enfoque de la léxico-métrica, que es una disciplina enmarcada en la lexicología y lexicografía, cuyo objetivo es conocer el léxico de una lengua, según las características estadísticas de las palabras (Valencia y Echeverría, 1999; Pérez, 2005; Valencia, 2011; Castillo Fadić, 2021a). De manera específica, el LD se aborda desde la disponibilidad léxica (DL). Esta última se refiere al

constructo teórico y metodológico por medio del cual se obtiene y analiza el LD (Hernández Muñoz y Tomé, 2017; Gómez Molina, 2021).

En virtud de lo anterior, puede afirmarse que el LD, por una parte, favorece la evaluación del nivel léxico de la lengua y, por otra parte, permite describir una parte importante de las competencias lingüísticas de los hablantes, fundamentado en una metodología ampliamente probada. Ergo, resulta pertinente resaltar el papel que juega la disponibilidad léxica en el análisis del vocabulario de estudiantes universitarios, desde una perspectiva sociolingüística, sobre campos semánticos generales, como: *Partes del cuerpo*, *Comidas y bebidas*, *La escuela: muebles y materiales*, *Juegos y diversiones*; pero también respecto a áreas nocionales próximas a las mallas curriculares de los programas académicos, como: *La lectura*, *El profesor*, *La educación*, *Habilidades y cualidades docentes*. Con la finalidad de comprender las aseveraciones previas y enmarcar la necesidad de una propuesta metodológica novedosa y, por ende, alternativa en la recogida de datos léxico, a continuación, se expone un breve recorrido histórico de los avances en los estudios léxico-estadísticos, para luego exhibir la evolución de las pesquisas en DL.

La léxico-estadística, como también se denomina a la léxico-métrica, tiene una larga tradición que puede rastrearse hasta finales del siglo XIX, cuando aún no se contaba con la ayuda de las ciencias computacionales y la informática, tan esenciales hoy para la recolección, almacenamiento, procesamiento y análisis de los corpus. Tal como se desprende del arqueo bibliográfico, el pionero de este campo disciplinar es Känding (1897), al que siguieron lexicógrafos como: Auber (1953) y Varlée (1934), en francés; Ayres (1915), Zipf (1946), en inglés; Josselson (1953), en ruso; Morales (1986), Sebastián, Martí y Carreira (2000), Castillo Fadić (2021a), en español; Mathy (1952) sobre el latín, entre otros.

J. W. Käding publicó a finales del siglo XIX *Häufigkeitswörterbuch der Deutschen Sprache* (1897). Se trata de la primera obra lexicográfica en la que se utiliza un criterio matemático para la selección léxica (en alemán). A esta, le siguieron: *Teacher's Word Book* de Thorndike (1921); *A graded Spanish Word Book*, de Milton Buchanan (1929), quien, además, planteó el primer índice matemático que combinaba frecuencia y rango, mediante la fórmula: $\left(\frac{f}{10} + r\right)$ (Buchanan, 1929: 10). George Vander Beke (1935) también utilizó la ecuación de Buchanan, en *French Words Book*. Asimismo, pueden enumerarse los léxicos reducidos de Aristizábal M. (1938) y Tharp (1939). Para finalizar, no puede dejarse afuera la investigación de Juilland y Chang-Rodríguez (1964), titulado *Frequency Dictionary of Spanish Words* (cf. Lara, 2006; Ávila Martín, 2010; Castillo Fadić, 2021a).

Este breve recorrido ha intentado trazar una línea cronológica en la confección de los diccionarios estadísticos, denominados así "No porque tengan como tema central la estadística ni porque esté dirigido a expertos matemáticos, sino porque se basa en cálculos estadísticos y presentan, en vez de las definiciones que solemos encontrar en la mayoría de los diccionarios, índices estadísticos" (Castillo Fadić, 2021a: 19). Gracias a dichas obras, los lexicógrafos han colaborado con innovaciones de índole metodológica, tales como: i) las fórmulas matemáticas, ii) las técnicas de construcción de los corpus y iii) el tratamiento de la homonimia y polisemia, por mencionar algunas. Paralelamente, este somero recorrido histórico pretende ilustrar las raíces de la DL. No obstante, debe tenerse en cuenta que la epistemología de la DL es distinta a la de los otros repertorios léxicos, tales como los de frecuencia léxica y léxicos básicos.

De manera puntual, los trabajos acerca del léxico disponible se iniciaron en Francia durante la segunda mitad del siglo XX, en el marco de la configuración de *Le français élémentaire* (1956) de Gougenheim, Michéa, Rivenc y Sauvageot. Este proyecto buscaba crear materiales didácticos para la enseñanza y aprendizaje del francés de forma más expedita, con lo que se pretendía fortalecer la expansión de la lengua gala en el mundo, y con ella, la cultura francesa (cf. Gougenheim, 1954, 1955; López Morales, 1995-1996, 1999; Ávila Muñoz y Villena, 2010, entre otros). En este contexto, los lexicógrafos franceses –con las contribuciones de Michéa– recurrieron a métodos fundamentados en la psicología experimental y la pedagogía para el diseño metodológico que guiaría sus trabajos. Desde entonces, la DL (línea de investigación emergente) comenzó a extenderse y a desarrollarse en diversas lenguas.

Entre los precursores puede contarse a Dimitrijević, quien en 1969 publicó *Lexical Availability. A New Aspect of the Lexical Availability of Secondary School Children*, obra que recogía el léxico disponible del inglés. Por su lado, Mackey (1971) abordó el léxico del francés de Canadá; en tanto, Njock (1978) aportó un recuento de léxico potencial en dos lenguas, gracias al análisis de los listados de palabras elaborados por niños bilingües francés-basaa. Igualmente, Bailey Victery (1971) enfocó su atención en el bilingüismo, por lo que analizó el LD de hablantes de inglés y español. En cuanto a López Morales (1973), el lingüista centró su atención en el español de Puerto Rico, siendo esta la investigación pionera en la que se analizaba únicamente la lengua de Cervantes. Así pues, los estudios que empezaron en el idioma galo pronto se extendieron a otros.

Empero no solo el interés por la DL se midió por el número de lenguas en las que eran replicados los trabajos de los lexicógrafos franceses, sino también por su contribución a otras disciplinas. Puesto que, si bien la DL nació en el marco de la lingüística aplicada, su metodología ha

motivado a pesquisidores de otros campos de las ciencias del lenguaje. Al respecto, pueden mencionarse las investigaciones de López Morales (1973, 1994a, 1999), Román (1985), Mena Osorio (1986), Justo Hernández (1987), Valencia y Echeverría (1999), Valencia (1994, 2010, 2011), Hernández Muñoz (2006), Ávila Muñoz y Villena (2010), entre otras, las cuales han abordado el caudal léxico potencial desde la psicolingüística, dialectología, sociolingüística, etc. En efecto, los trabajos realizados en la línea de la psicolingüística han apuntado *grosso modo* a explicar la configuración del lexicón, la adquisición y ampliación del vocabulario, entre otras cuestiones lingüísticas. Por su parte, las pesquisas dialectológicas han intentado determinar las relaciones interdialectales de las palabras y establecer los usos y vitalidad del léxico en distintas regiones, dibujando isoglosas. En tanto que los estudios llevados a cabo desde la sociolingüística han buscado conocer la realidad lingüística y la variación del vocabulario de las distintas comunidades de habla. Esto ha tenido la finalidad de establecer la norma del léxico potencial de las sintopías exploradas. Una mirada crítica al arqueo bibliográfico apunta a que, aparentemente, el grueso de este tipo de pesquisas se ha realizado desde la sociolingüística. Entre algunas de las razones que podrían explicar este panorama se encuentra que los métodos de ambas disciplinas se complementan.

Respecto de algunas líneas de investigación de la lingüística cognitiva, a veces a caballo con la psicolingüística, han encontrado en la DL una herramienta efectiva para abordar estudios acerca de las asociaciones entre las palabras y los núcleos prototípicos, entre otros aspectos. En este sentido, la disponibilidad léxica puede concebirse como un campo lingüístico interdisciplinar que colabora con el entendimiento del lenguaje desde el análisis del caudal de vocabulario de una determinada comunidad de habla. Particularmente, pueden referirse los trabajos que han fundamentado sus análisis en la teoría de grafos. Al respecto, Henríquez *et al*. (2016) describieron los distintos mecanismos cognitivos que se activaban en la producción y asociación del léxico disponible. Los autores analizaron los datos referidos al eje temático *Cuerpo humano*, según dos grandes mecanismos, los cuales contaban con subtipos, a saber: 1) Semánticos-Cognitivos (*Categorización*, *Hiperonimia e hiponimia*, *Sinonimia*, *Antonimia y opuestos*) y 2) Lingüísticos-Formales (*Colocaciones*, *Composición sintáctica*, *Asociación morfológica*, *Asociación fonética*). Sus resultados indicaron que el mayor porcentaje de asociaciones de palabras eran del primer tipo (Semánticos-Cognitivos), lo que podría explicarse por la propia naturaleza de la prueba de DL.

Una vez sondeado *grosso modo* el grado de interdisciplinariedad de la DL, en los próximos párrafos se pondrá el foco en las pesquisas llevadas a cabo en español. El arqueo bibliográfico exhibe lo fructífera que ha sido la DL en el mundo hispánico (López Morales y Trigo Ibáñez, 2019), donde

se encuentran trabajos tanto en la península, como en las Canarias, el Caribe, el centro y el sur de América. En un primer momento, los trabajos de DL en la lengua castellana se centraron en estudiantes de secundaria, generalmente desde un enfoque sociolingüístico. Sin embargo, –como indican López Morales (1995-1996), Ávila Muñoz y Villena (2010) y Hernández Muñoz y Tomé (2017), entre otros– pronto brotaron también exploraciones orientadas hacia la dialectología, psicolingüística y etnografía (cf. Galloso, 1998; Gómez Molina y Gómez Devís, 2004; Mahecha y Mateus Ferro, 2017; Trigo Ibáñez, Romero y Santos Díaz, 2019; Martínez-Lara, 2021, 2023, entre otras). Asimismo, este campo lexicológico se amplió hacia otras realidades lingüísticas, como el léxico de los niños (Mesa Canales, 1989; Gómez Devís, 2019; Gómez Devís y Cepeda, 2022; Escudero, Santos Díaz y Trigo Ibáñez, 2022); adultos mayores (Echeverría y Urrutia, 2004; Urzúa, 2018; Rojas Zepeda, 2020); comunidades de hablas estratificadas sociolingüísticamente (López Morales, 1999; Valencia y Echeverría, 1999; Ávila Muñoz y Villena, 2010) e incluso hacia el área de la literacidad en salud (Castillo Fadić y Pino Castillo, 2020; Castillo Fadić y Santos Díaz, 2021).

Las investigaciones en español como lengua materna impulsaron los estudios de DL de español como lengua extranjera (cf. Lin, 2012; Serfati, 2016; Mendoza Puertas, 2018; Herranz y Marcos, 2019; Aabidi, 2020; Santos Díaz, Trigo Ibáñez y Romero, 2020; Zhou, 2021). Igualmente, ha habido un fomento por analizar el caudal léxico de los hablantes del español que han cursado estudios en lenguas extranjeras (cf. Ferreira *et al.*, 2019; Santos Díaz, 2020; Santos Díaz y Juárez, 2022). A su vez, empezaron a surgir nuevas aplicaciones de la DL, como la creación y el análisis de corpus cacográficos y descripción de procesos neológicos (cf. García, 2011; Paredes García, 2012a, 2018; Paredes y Gallego, 2019; Trigo Ibáñez *et al*., 2018).

De igual modo, la revisión bibliográfica indica que ha habido un interés creciente por determinar el léxico disponible de las comunidades discursivas especializadas. Al respecto, pueden señalarse las pesquisas llevadas a cabo con estudiantes universitarios de áreas como: Educación, Humanidades, Salud, Ingeniería, Veterinaria, Comunicación Social, entre otras (cf. Gómez y Guerra, 2004, 2005; Ávila Muñoz, 2007; Fasce *et al.*, 2009; Ferreira, Salcedo y Leo, 2014; Rojas, Zambrano y Salcedo 2017; Blanco *et al.*, 2020; Herranz, 2020; Santos Díaz, 2020; Marcos-Calvo y Herranz, 2021, Fregoso-Peralta y Aguilar-González, 2022). En estos estudios se ha buscado conocer el LD de los futuros profesionales en ejes temáticos concernientes a las especialidades de las respectivas mallas curriculares. En algunos casos, se han contrastado los lexicones de los universitarios con los del profesorado o con el de profesionales en ejercicio (cf. Urzúa, Sáez y Echeverría, 2006; Fasce *et al.*, 2009; Kloss y Quintanilla, 2022).

En virtud de que la DL representa una excelente herramienta para conocer el caudal léxico del alumnado, Santos Díaz y Juárez (2022: 267) han señalado que la relevancia que ha tenido esta disciplina en la lingüística aplicada ha llevado a los lexicógrafos a poner el foco de atención también en "la evaluación de la competencia léxica del futuro profesorado", con el fin de conocer la incidencia de la formación universitaria en el lexicón de los discentes (cf. Herranz, 2018; 2020; Castillo Fadić y Sologuren, 2020; Cerda *et al.*, 2017; Valenzuela *et al.*, 2018; Quintanilla y Salcedo, 2019; Santos Díaz, 2020; Marcos-Calvo y Herranz, 2021; Martínez-Lara, 2021; Zambrano, 2021). Así pues, puede señalarse que la DL no solo ha contribuido con la descripción del léxico disponible en relación con factores sociológicos, dialectales, etnográficos y psicolingüísticos, sino también con categorías ligadas a la metacognición de las comunidades discursivas. En este sentido, resulta relevante testear de qué manera los estudiantes universitarios aportan al caudal léxico de un campo semántico particular en relación con la carrera que cursan.

No obstante, en el contexto de la educación superior, aún son escasos los estudios en los que se comparen los caudales léxicos de universitarios de distintas especialidades. A propósito de esto, puede mencionarse la investigación de Guerra y Gómez (2004), quienes analizaron el LD de alumnos regulares de Periodismo, Comunicación Audiovisual y Derecho, en cuanto a tres ejes temáticos, a saber: *Prensa*, *Radio* y *Televisión*, los cuales estaban entroncados con la malla curricular de las dos primeras carreras. Estos lexicógrafos buscaban conocer, por un lado, el vocabulario utilizado por los discentes de Periodismo y Comunicación Audiovisual con la finalidad de identificar el léxico que debía enseñárseles a los extranjeros que cursaban dichos programas en España. Por otro lado, querían determinar las semejanzas y diferencias del léxico de alumnos de áreas disímiles. Entre las conclusiones, Guerra y Gómez (2004) indicaron que, si bien el alumnado de las tres especialidades comparte un vocabulario común indistintamente de las materias en las que se especializan, existen disimilitudes cualitativas que dan cuenta de las diferencias disciplinares. En efecto, se apreció un vocabulario más coincidente entre los datos de Periodismo y Comunicación Audiovisual, con palabras de uso especializado. Este se distinguió del repertorio de los de Derechos, en los que se observó un mayor número de palabras comunes. En este sentido, podría argumentarse que los estudiantes universitarios de una carrera X –piénsese, por ejemplo, en Letras Hispánicas– aportarán no solo más palabras, sino también un vocabulario más específico, en un campo semántico particular, como *La lectura*, ya que este se vincula estrechamente con el programa que cursan; mientras que un grupo de estudiantes de una carrera Y –por ejemplo, Ingeniería–, en el mismo eje temático, posiblemente, reportará un léxico de conocimiento general.

En esta misma línea se encuentra el trabajo de Blanco *et al.* (2020), quienes compararon los lexicones de estudiantes de las facultades de Educación y Humanidades y Artes de la Universidad de Concepción, Chile, con el objetivo de conocer las palabras más disponibles relacionadas con las emociones, específicamente: *Rabia*, *Sorpresa*, *Amor*, *Alegría*, *Miedo*, *Tristeza* y *Asco*. Los resultados de los promedios de palabras e índices de cohesión apuntaron a que las distinciones entre ambos grupos eran mínimas. No obstante, a diferencia del caso de Guerra y Gómez (2004), los analistas chilenos no tomaron la variable carrera como un factor extralingüístico que pudiera explicar las convergencias y divergencias del léxico de las emociones. En consideración a lo anterior, podría suponerse que las exploraciones de los campos nocionales de índole general, como las emociones, *Partes del cuerpo* o *Comidas y bebidas*, entre otros, aportan eficazmente a los estudios contrastivos con comunidades de habla y discursivas (análisis intergrupal) distintas. Además, estas contribuyen con la validación de la metodología aplicada. Sin embargo, los trabajos basados en campos nocionales generales podrían resultar poco efectivos en análisis intragrupales que buscan conocer las (di)similitudes entre comunidades disciplinares. En este caso debería recurrirse a ejes temáticos ligados a las áreas de conocimiento, como en las pesquisas de Herranz (2018, 2020), Martínez-Lara (2021) y Fregoso-Peralta y Aguilar-González (2022).

En resumen, aunque aparentemente son escasas las reflexiones sobre LD de universitarios de distintos programas, está comenzando a trazarse un camino en esa dirección, como se ha visto en los párrafos previos. Ahora bien, salvando estos progresos referidos a la selección de la población de estudio, también es necesario pensar en nuevos métodos que permitan llegar a distintos sociolectos y comunidades discursivas sin contar con el requisito privativo de la presencialidad y el formato en papel. Este planteamiento se fundamenta en el hecho de que, en un mundo en el que cada día las distancias se aminoran gracias a la internet –lo que facilita las conexiones y las transferencias de datos entre los usuarios–, las aplicaciones tecnológicas vienen a jugar un rol crucial en la actualización de los métodos de investigación. En consideración a lo anterior, deben destacarse los avances en la lingüística de corpus, ya que esta ha sabido aprovechar los recursos computacionales, no solamente para construir corpus de millones de palabras, sino también para el procesamiento y análisis de los datos (Martínez-Lara, 2016). También deben resaltarse algunas pesquisas llevadas a cabo desde la pragmática, especialmente desde la cortesía lingüística, en las que los datos han provenido de distintas plataformas electrónicas (cf. Yus, 2001, 2010; Dorta, 2008; Fröhlich y Lux, 2008; Peng y Moreno, 2021, entre otros).

Sin embargo, no solo las disciplinas lingüísticas han considerado dar un paso al frente en el uso de la internet y las ciencias de la computación para logar sus objetivos, sino otras áreas, como Educación. En esta se ha llegado a recurrir al uso de plataformas didácticas, las cuales funcionan como complementos en el proceso de enseñanza-aprendizaje. De manera excepcional, debe mencionarse el escenario de la crisis originada por la pandemia del SARS-CoV-2 que llevó a replantear la manera cómo impartir clases y, a su vez, prevenir los contagios y nuevos brotes de la enfermedad. Una de las soluciones fue apelar al uso de las aplicaciones digitales con la finalidad de darle curso a los objetivos pedagógicos en los diferentes programas formativos. Con esta medida, se esperaba evitar el atraso y paralización de la enseñanza formal. En este contexto, lo investigadores realizaron estudios en los que se evaluaron los alcances y limitaciones de la formación asincrónica y el uso de internet con fines didácticos (cf. Del Mastro y Albujar, 2021; Fainholc, 2021, entre otros). Los resultados de algunas de las investigaciones han exhibido las ventajas y desventajas de trabajar de la mano con las nuevas tecnologías, lo que ha despertado las reflexiones en torno a los métodos didácticos. Con este panoramas, no parece ser extraordinario plantear la creación de herramientas digitales con las que puedan aplicarse pruebas de disponibilidad léxica y, por ende, recolectar las palabras potenciales evocadas por estas. De hecho, Blanco *et al.* (2020) ya expone someramente un prototipo computacional creado por el profesor Pedro Salcedo por medio del cual se recopila LD.

En consideración al punto anterior: uso de aplicaciones computacionales en la toma de LD, la revisión bibliográfica direccionó la atención hacía técnicas alternativas a las que se han recurrido en la recogida de los corpus de DL. Una de ellas ha sido la prueba oral. En esta, en vez de utilizarse cuadernillos de papel, se graban las respuestas que van evocando los hablantes por cada centro de interés (cf. Murillo y Sánchez, 2006; Hernández Muñoz, 2004; Tomé, 2015; Herreros, 2016; Zhou, 2021). Acerca de este tipo de examen, Zhou (2021) propone el formato de las pruebas (escrito y oral) como una variable independiente de su estudio de LD, con lo cual buscaba contrastar los lexicones recolectados manual y oralmente, con la finalidad de evaluar ambas técnicas. Los resultados de los promedios de palabras por esferas semánticas indicaron que los hablantes chinos produjeron más piezas léxicas en modalidad oral que en papel, siendo *Comidas y bebidas*, *Economía e industria* y *Escuela y universidad* los actualizadores más productivos. La autora concluye que esta variable resultó estadísticamente significativa, por lo que se asume la influencia lingüística sobre ella (Zhou, 2021: 343). Asimismo, ha habido excursiones en otras tácticas para la conformación de corpus de palabras potenciales. Por ejemplo, Ferreira *et al.* (2019), quienes evaluaron los factores asociados con el LD en inglés como lengua extranjera (L2) de futuros profesores de inglés, recurrieron a la recolección de

lexicones gracias a computadoras; las pruebas se realizaron *in situ*, en documentos Word, no a través de un programa electrónico particular; tampoco se buscaba determinar la idoneidad de la utilización de ordenadores.

Igualmente, pueden mencionarse los trabajos de Ávila Muñoz, Santos Díaz y Trigo Ibáñez (2020); Blanco *et al.* (2020) y Martínez-Lara (2023), en los que se recurrieron a artilugios digitales. En el primero, se tenía el objetivo de determinar las percepciones de la realidad, desde una perspectiva cognitiva experiencial, reportadas por estudiantes universitarios españoles en el contexto de la cuarentena por el covid-19. Los campos semánticos analizados fueron *Pandemia, Confinamiento, Futuro* y *Política*. Para esto, los investigadores tomaron en cuenta *grosso modo* la metodología de la disponibilidad léxica, la cual adaptaron a sus objetivos (Ávila Muñoz *et al.*, 2020: 86). Los listados de palabras se obtuvieron mediante una encuesta creada en la plataforma de Google Forms. Uno de los inconvenientes de esta herramienta computacional compete a que no es posible establecer el tiempo de respuesta por área nocional, lo que se contrapone a los planteamientos epistémicos de la DL. En esta misma línea, Martínez-Lara (2023), siguiendo a grandes rasgos la iniciativa de los lingüistas españoles, analizó el LD de un grupo de 227 venezolanos con el objetivo de describir las convergencias y divergencias dialectales sobre las esferas semánticas *Comidas, Bebidas* y *Pandemia*. Para esto, los datos léxico se tomaron a través de Google Forms y fueron analizados, según la metodología de DL. Si bien el área nocional *Pandemia* se halla tanto en la pesquisa española como en la venezolana, los resultados no pudieron compararse, porque se utilizaron cálculos e índices diferentes. En el trabajo de Ávila *et al.* (2020) se aplicó el índice de centralidad (Ávila-Muñoz y Villena-Ponsoda, 2010), mientras que en el de Martínez-Lara (2023), el índice de disponibilidad léxica (López Chávez y Strassburger Frías, 1987). No obstante, ambos estudios han contribuido con la reflexión acerca del uso de técnicas computacionales mediadas por internet para la recogida del LD.

En la esfera chilena, como se indicó previamente, Blanco *et al.* (2020), observaron el LD de estudiantes de la Universidad de Concepción, según la teoría de grafos. Los lexicones analizados se recogieron a través de una aplicación electrónica que, a diferencia de Google Forms, sí medía el tiempo de respuesta de los sujetos (2 minutos por campo nocional). No obstante, Blanco *et al.* (2020) no ofrecen detalles acerca de la configuración de la herramienta digital ni la manera cómo llegaron a los sujetos.

En síntesis, a pesar del surgimiento de nuevas técnicas y métodos de recopilación del léxico disponible, basadas en las ciencias computacionales, no se han realizado estudios de DL en los que se contrasten datos recolectados en cuadernillos de papel o método tradicional y lexicones tomados vía

plataforma electrónica, mucho menos se han evaluado dichas aplicaciones. Aunado a esto, los datos y resultados con los que se cuentan actualmente, como los de Ferreira *et al.* (2019), Ávila Muñoz *et al.* (2020), Blanco *et al.* (2020) y Martínez-Lara (2023), no pueden compararse, debido a la disparidad de los métodos, teorías y actualizadores. En este sentido, surge la necesidad de ahondar en los alcances y limitaciones que podrían tener los recursos digitales para la recogida de LD, con lo cual se avanzaría en las propuestas metodológicas de este campo disciplinar.

En resumen, la sociolingüística ha resultado ser una disciplina enriquecedora de los estudios de disponibilidad léxica (Samper Padilla, 2021), ya que permite determinar el estado real del vocabulario de un grupo social particular, por ejemplo: escolares chilenos (Valencia y Echeverría, 1999) o navarros (Jiménez, 2019); hablantes malagueños (Ávila Muñoz y Villena, 2010), entre otros. Asimismo, contribuye con la explicación de la variación diastrática del léxico, según diversos factores, como: sexo (Prado y Galloso, 2008), nivel sociocultural (Arnal, 2008), tipo de centro educativo (Trigo Ibáñez y Santos Díaz, 2021), entre otros. Respecto a los centros de interés, los antecedentes muestran que –al desarrollar análisis contrastivos entre distintas sintopías– debiesen utilizarse campos nocionales de conocimiento general, compartido por toda la comunidad, como: *Partes del cuerpo*, *Juegos y distracciones*, *La ropa*, etc. Empero, al conjugar grupos de comunidades científicas o disciplinares disímiles, sería recomendable testear actualizadores relacionados con las áreas de conocimiento exploradas. En cuanto a las técnicas de la DL, surge la necesidad revisar los métodos de recogida de LD a la par de los avances tecnológicos, porque estos motivan la migración hacia modelos computacionales que robustezcan y actualicen las metodologías.

A partir de los puntos expuestos previamente, esta tesis doctoral compete a un estudio sociolingüístico del léxico disponible –respecto de los campos nocionales: *La lectura*, *El profesor*, *La educación*, *Juegos y distracciones*, *La escuela: muebles y materiales*, *Habilidades y cualidades docentes*, *Partes del cuerpo* y *Comidas y bebidas*, recolectado a través de pruebas en dos formatos: papel y digital– de estudiantes regulares de las carreras de Educación Básica y Letras Hispánicas de la Pontificia Universidad Católica de Chile.

En este contexto, la investigación se inscribe en el campo de la lexicología y la lexicografía, específicamente en la línea de la lexicometría, a fuer de que atañe al nivel léxico de la lengua. Respecto al alcance temporal, se ajusta a los parámetros sincrónico, puesto que los datos incumben a un único periodo, aunque los análisis comparativos de la variable *Año de curso* podrían relacionarse con el tiempo aparente. Pero no se ha establecido una perspectiva diacrónica en el estudio, además, el corpus no es adecuado para tal fin. Asimismo, este trabajo se enfoca diatópicamente en datos recogidos

solamente en la ciudad de Santiago de Chile, donde se ubica la población objeto de estudio. En cuanto a la profundidad y el carácter, la pesquisa es de tipo mixto, ya que, si bien el grueso de los análisis es cuantitativo –descriptivo e inferencial–, también se han realizados exploraciones cualitativas del léxico disponible. A su vez, a partir de la descripción de los datos, se ha llegado a la interpretación y explicación de estos, para lo cual se ha utilizado fuentes primarias, recolectadas, organizadas y procesadas por el mismo investigador responsable. Por último, si bien los léxicos disponibles constituyen sustentos inestimables para la enseñanza-aprendizaje de lengua, por lo que suelen abordarse desde la lingüística aplicada, esta investigación monográfica ha sido enfocada, en principio, hacia una perspectiva pura del vocabulario. Entonces, a partir de esta delimitación, se plantean las siguientes preguntas de investigación:

- ¿Cuánto y cuál es el léxico disponible de los estudiantes universitarios de las carreras de Educación Básica y Letras Hispánica de la Pontificia Universidad Católica de Chile, sobre los centros de interés: *La lectura*, *El profesor*, *La educación*, *Juegos y distracciones*, *La escuela: muebles y materiales*, *Habilidades y cualidades docentes*, *Partes del cuerpo* y *Comidas y bebidas*?
- ¿Cuáles son las convergencias y divergencias, cuantitativas y cualitativas, del léxico disponible respecto al programa de titulación de los participantes (Educación Básica y Letras Hispánicas) y del formato de las pruebas (papel y digital)?
- ¿Las variables extralingüísticas *Sexo*, *Carrera*, *Año o nivel de curso*, *Formato de pruebas*, *Cantidad de libros* y *Frecuencia de lectura* inciden en el caudal léxico de los participantes?
- En relación con las redes asociativas ¿cómo se organizan las palabras más disponibles del centro de interés más productivo, rico y compacto del presente estudio?
- En cuanto a los factores *Carrera* y *Formato de pruebas* ¿cuáles son las semejanzas y diferencias de las redes asociativas del CI más productivo y cohesionado de los corpus?

Con base en este conjunto de preguntas, el objetivo general de esta tesis doctoral es conocer, cuantitativa y cualitativamente, el léxico disponible de estudiantes universitarios de Educación Básica y Letras Hispánicas de la Pontificia Universidad Católica de Chile –recolectado mediante instrumentos en papel y digital–, referido a los centros de interés: *La lectura*, *El profesor*, *La educación*, *Juegos y distracciones*, *La escuela: muebles y materiales*, *Habilidades y cualidades docentes*, *Partes del cuerpo* y *Comidas y bebidas*, con la finalidad determinar las convergencias y

divergencias del caudal léxico de los grupos bajo análisis. A partir de este, se desprenden los siguientes objetivos específicos:

1. Determinar los índices generales y específicos de disponibilidad léxica de las palabras producidas por estudiantes de Educación Básica y Letras Hispánicas de la UC acerca de los centros de interés: *La lectura*, *El profesor*, *La educación*, *Juegos y distracciones*, *La escuela: muebles y materiales*, *Habilidades y cualidades docentes*, *Partes del cuerpo* y *Comidas y bebidas*.
2. Conocer las palabras más disponibles de los grupos de estudiantes universitarios analizados sobre los centros de interés planteados en la presente investigación.
3. Comparar los vocabularios potenciales de los sujetos analizados, con el fin de determinar las convergencias y divergencias a tenor de las categorías: *Sexo*, *Carrera*, *Año de curso*, *Formato de pruebas*, *Cantidad de libros leídos* y *Frecuencia de lectura*.
4. Evaluar el instrumento en formato electrónico de recolección de datos léxicos, a partir de los análisis contrastivos de los datos.
5. Determinar la incidencia de las variables sociológicas (*Sexo*, *Carrera*, *Año de curso*, *Formato de prueba*, *Cantidad de libros leídos* y *Frecuencia de lectura*) en el caudal léxico de los alumnos de Educación Básica y Letras Hispánicas.
6. Describir y explicar la conformación de las redes asociativas del centro de interés más productivo y cohesionado del estudio.
7. Respecto al actualizador más productivo, rico y cohesionados, comparar las redes asociativas en relación con el *Formato de pruebas* (en papel y digital).

En este punto, debe señalarse que se decidió analizar a través de grafos los datos del CI con el mayor promedio de palabras e índice de cohesión, debido a que estos criterios permiten observar con más detalles los fenómenos de asociación semántica de las palabras disponibles (cf. Mahecha *et al.*, 2017; Mateus-Ferro *et al.*, 2018 y Blanco *et al.*, 2020).

Contar con un repertorio del léxico disponible de una población tiene la ventaja de exhibir una parte de la lengua viva, puesto que muestra el catastro de la realidad lingüística de una comunidad. En el caso de cursantes de programas universitarios –particularmente los de las áreas de Educación y Letras–, si bien es pertinente identificar la adquisición de nuevos términos, sobre todo aquellos ligados a las respectivas mallas curriculares, también resulta relevante establecer su productividad léxica. Esto debido a que es menester que los futuros formadores e investigadores del lenguaje y la pedagogía manejen un vocabulario amplio y acorde a sus roles profesionales (cf. Valencia y Echeverría, 1999;

Santos Díaz, 2020). Asimismo, en un mundo cada vez más exigente, en el que las telecomunicaciones acercan los conocimientos a un mayor número de personas, los especialistas en didáctica e investigación en las ciencias del lenguaje deben estar cada vez más cualificados, con un componente lingüístico sólido y pertinente.

Acorde con los planteamientos previos, desde una perspectiva sociolingüística, los diccionarios de léxico disponible brindan una panorámica acerca de la vitalidad, variabilidad y productividad de las palabras pertenecientes a un campo nocional específico (*El campo*, *La ciudad*, *La ropa*, etc.) que comparte un colectivo. A través de este tipo de investigación, puede conocerse el caudal léxico que más activa una comunidad discursiva (Bolívar, 2013) en relación con los componentes académicos que ha adquirido, ya sea durante el proceso de aprendizaje formal, ya sea mediante consumos culturales, como la lectura (Ávila Muñoz, 2007). Así pues, esta tesis doctoral contribuye con un diccionario estadístico que incluye campos nocionales novedosos (*La lectura*, *El profesor* y *Habilidades y cualidades docentes*), el aplicado por Herranz (2020): *La educación*; y cuatro de los dieciséis actualizadores del PPHLD: *Juegos y distracciones*, *La escuela: muebles y materiales*, *Partes del cuerpo* y *Comidas y bebidas*. Asimismo, este trabajo intenta aportar conocimiento sobre la relación entre el léxico potencial y los factores extralingüísticos: *Sexo*, *Carrera*, *Año de curso*, *Formato de prueba*, *Cantidad de libros leídos* y *Frecuencia de lectura optativa*. Por último, este estudio supone, un aporte al avance, innovación y fortalecimiento de la metodología del campo de la disponibilidad léxica, al proponer y evaluar una forma alternativa –acorde con la era digital– para la recogida de las listas de palabras disponibles.

Después de esta sección introductoria, en el capítulo 1, *Marco teórico*, reseña el arqueo bibliográfico. En la primera sección se ubican los estudios de disponibilidad respecto a los marcos metodológicos utilizados en la elaboración de vocabularios reducidos (lógico y estadístico); seguidamente, se expone la diferencia entre léxico disponible y disponibilidad léxica. En la tercera sección se definen los conceptos fundamentales de DL. En la cuarta se expone un recorrido por el desarrollo de la DL en el marco del Francés fundamental; mientras que en la quinta sección se reseñan los estudios precursores de DL, pasando a la sexta parte, donde se detallan algunos aspectos del Proyecto Panhispánico de Léxico Disponible; y, finalmente, se explica la fórmula para calcular el índice de disponibilidad léxica.

El capítulo 2, *Metodología*, está destinado a explicar los distintos pasos –fundamentado en el marco teórico y los antecedentes– que se siguieron para la realización de este trabajo. Primeramente, se detalla el tipo de investigación en el que se encuadra la tesis; seguidamente, se indican los procesos

de selección de los informantes; luego, se informa acerca del tamaño muestral, basado en los cálculos de potencia estadística; y se muestran las características de los participantes. En la quinta sección se describen los instrumentos: i) cuestionario sociológico, ii) test de disponibilidad léxica y iii) encuesta de prácticas lectoras; mientras que en el sexto epígrafe se puntualizan los métodos y técnicas de recolección de datos, con énfasis en la descripción de la página web. En la séptima parte se muestran los procesos de edición de los materiales; seguidamente, se plantea la codificación y, por último, se explica el procesamiento del corpus.

Los *Análisis cuantitativos generales* se plasman en el capítulo 3, que empieza con la exposición de los resultados concernientes al número de palabras. Esta se desarrolla desde el Número de palabras del corpus general, en el que se recogen las respuestas de los 264 informantes; seguidamente, el número de lexías del subcorpus 1, es decir, el compuesto por las muestras recolectadas en papel; y, finalmente, el cómputo de las unidades léxicas del subcorpus 2, en el que se hallan los datos de los alumnos de Letras Hispánicas que realizaron los test en papel y digital. La segunda sección se refiere al análisis del número de vocablos; mientras que en la tercera sección se indican los índices de dispersión; luego, se presentan los promedios de palabras; y finalmente, los resultados sobre los vocablos más disponibles por centro de interés de las tres muestras cierran el capítulo. Todos estos subapartados siguen la misma estructura indicada para el número de palabras.

En el capítulo 4, *Análisis sociolingüístico*, se reseñan los resultados descriptivos de las variables independientes del estudio, a saber: *sexo*, *carrera*, *año de curso*, *formato de prueba*, *cantidad de libros leídos* y *frecuencia de lectura optativa*. Una vez conocido los atributos cuantitativos de la distribución de los informantes por cada una de las variables, se prosiguió a realizar los análisis bivariados con t de Student, para las categorías dicotómicas, y Anova, para las politómicas. Estos análisis se llevaron a cabo gracias al programa SPSS®.

Los *Análisis cualitativos* se localizan en el quinto capítulo, en el que se muestran las convergencias y divergencias del LD de los encuestados, a tenor de las variables socioeducativas: *sexo*, *carrera*, *año de curso*, *formato de prueba* y *frecuencia de lectura optativa*, con la intención de intentar explicar las relaciones entre el vocabulario y los rasgos extralingüísticos de los participantes. Igualmente, se describieron los grafos referentes al centro de interés *partes del cuerpo*; se exploró este por ser el más productivo y compacto de los ocho de la tesis. Asimismo, se compararon los grafos de cada muestra.

En el capítulo 6, *Conclusiones*, se recogen las apreciaciones finales que se desprendieron de los resultados y análisis de los datos léxicos. Estas se van reseñando en función de los objetivos

específicos planteados. Y, finalmente, a este último apartado, le siguen las referencias bibliográficas, los anexos y los diccionarios de LD de las muestras en papel de Educación Básica y Letras Hispánicas y del corpus digital de LH.

Capítulo 1. Marco teórico

1. Consideraciones previas

Los estudios léxico-métricos se enfocan en la descripción de las propiedades y tendencias estadísticas de las palabras, asimismo contribuye con la categorización de las unidades léxicas (Romero-Pérez *et al.*, 2018). Para esto, los lexicógrafos analizan, gracias a fórmulas matemáticas, corpus –tanto orales como escritos– de comunidades de habla, sociolectos y comunidades discursivas especializadas, con la finalidad de conocer los patrones de organización de las piezas léxicas en el lexicón. Al respecto, Vázquez-Cano *et al.* (2015: 88-89) afirman que:

> Entendemos, pues, el análisis estadístico de un texto como aquellos procedimientos que, mediante el cómputo de las ocurrencias de una o varias unidades verbales básicas, permiten realizar, a partir de los resultados obtenidos, algún tipo de cálculo, "de reorganizaciones formales de una secuencia textual y análisis estadísticos con el vocabulario resultante de una segmentación". Dichos análisis se conocen generalmente con el nombre de "Lexicometría".

En este orden de ideas, los estudios estadísticos del léxico han contribuido no solo con la exploración de grandes corpus textuales, sino también con el reconocimiento y taxonomía de las palabras, la creación de fórmulas matemáticas (complejas) para la detección y selección del vocabulario, la definición de los temas y campos semánticos de los materiales analizados, la relación interdisciplinar de corrientes lingüísticas, estadísticas y computacionales. En este contexto, se enmarcan los estudios de disponibilidad léxica, tema central de esta tesis doctoral. En virtud de lo anterior, los objetivos de este capítulo son: 1) detallar a grandes rasgos los métodos lógicos y estadísticos, utilizados generalmente en los estudios del léxico, particularmente, en la elaboración de vocabularios reducidos: léxico frecuente, básico y disponible. 2) Conocer la distinción entre léxico disponible y disponibilidad léxica, así como los conceptos fundamentales que sustentan las investigaciones de disponibilidad. 3) Reseñar el contexto inicial de los estudios de DL, desde el proyecto del Francés elemental hasta los trabajos precursores; 4) pormenorizar los hitos de las investigaciones de LD realizadas en América y el surgimiento del Proyecto Panhispánico de Léxico Disponible. Por último, 5) conocer la fórmula estadística aplicada en los cálculos de los índices de disponibilidad léxica.

En relación con los objetivos arriba planteados, este capítulo se organiza en ocho partes. En la primera, se presenta una caracterización acerca de los métodos lógico y estadístico, a los que se recurren para la construcción de vocabularios reducidos. En esta se exponen algunos ejemplos de diccionarios de frecuencia y léxico básico. En la segunda, se describen las distinciones entre los términos léxico disponible y disponibilidad léxica, con el fin de aclarar la terminología de este trabajo.

En la tercera parte, se explican los conceptos básicos en los estudios de DL, tales como: lexicología y lexicografía; palabra y vocablo; palabra temática y atemática; centro de interés, entre otros. En la cuarta parte, se recorre el contexto en el que surgió la disponibilidad léxica, como disciplina y herramienta léxico-métrica, desde los pormenores del proyecto del Francés elemental hasta sus resultados. En la quinta parte, se reseñan algunos de los primeros trabajos de disponibilidad léxica llevados a cabo a partir del *francés fundamental*. En la sexta, se enfoca la atención en conocer algunos de los primeros estudios de DL desarrollados en lengua española, especialmente en Puerto Rico, México, Chile y España. En la séptima, se comenta la creación y objetivos del Proyecto Panhispánico de Léxico Disponible. Por último, en la octava parte, se expone el camino recorrido en la elaboración de la fórmula para el cálculo del índice de disponibilidad léxica.

1.1. Enfoques metodológicos utilizados en la elaboración de vocabularios reducidos

Mateo García (1997), parafraseando a Mena Osorio (1986), señala que dentro de la lingüística hay dos posturas o corrientes que guían la construcción de diccionarios estadísticos, a saber: el método lógico y el estadístico. La primera se enfoca en la observación rigurosa de los datos con el objetivo de detectar patrones –generalmente definidos por razones de índole cualitativo, como las semejanzas y diferencias semánticas de un grupo de palabras– que permitan describir las unidades léxicas bajo análisis. A partir de lo anterior, los casos se agrupan y organizan sistemáticamente en las categorías que emergen de los materiales del corpus, según los criterios dictados por los investigadores.

El método lógico suele fundamentarse en los procedimientos de inducción y deducción (Herrera, Martínez y Amengual, 2011). Con esta premisa, puede acotarse que los investigadores parten de una (o un conjunto de) hipótesis que se confirma o rechaza por medio de los análisis cualitativos aplicados. Sin embargo, los especialistas también suelen optar por empezar a procesar primero los datos, con lo cual intentan identificar y, por ende, extraer las categorías explicativas del fenómeno observado. En este tipo de empresas, pueden llegar a utilizarse, en mayor o menor grado, criterios particulares, basados en la percepción que se tiene de los hechos (lingüísticos). No obstante, esta manera de construir las matrices de análisis puede conllevar a la subjetividad, por lo que los resultados podrían quedar comprometidos. Asimismo, podría darse el caso de que las conclusiones sean parciales. Entonces, la escogencia de los lemas de un vocabulario de lengua base realizada por medio de este método podría alejarse de la realidad lingüística que se ha querido describir (Mena Osorio, 1986: 20-21, citada por Marco 1997: 41-42).

Dentro de los estudios lexicológicos abordados desde el modelo lógico, puede reseñarse el trabajo de Ogden y Richards, quienes en 1928 publicaron *Basic English*[1], cuyo objetivo era proveer de un léxico reducido de la lengua anglo que fuese sencillo de comprender y aprender por hablantes de otras lenguas. Se trataría de un vocabulario instrumental –integrado únicamente por los elementos léxicos considerados (por los directores del proyecto) realmente importantes– que permitiera una comunicación básica y efectiva entre las personas. Para lograrlo, los autores seleccionaron las palabras que más se repetían en las definiciones de los diccionarios. Igualmente, fijaron un filtro que contemplaba los rasgos de: i) emotividad, ii) matiz literario y estilístico y iii) especialidad. Al finalizar, el vocabulario quedó constituido por 850 vocablos (600 sustantivos, 150 adjetivos y 100 palabras estructurales: operadores, pronombres y otros), a este los autores sumaron una lista complementaria de 1143 voces, con la cual el texto llegó a contemplar un total de 1993 piezas léxicas (Marco, 1997: 42; Sánchez-Saus, 2011: 51).

En la misma línea de Ogden y Richards (1928), Michael West publicó en 1941 *Vocabulario de definiciones*, cuyo objetivo era elaborar un léxico reducido del inglés que fuera significativo para la comunicación general. Para lo cual, el investigador se basó en la selección de las palabras más recurrentes en las definiciones de los diccionarios. Empero, a diferencia de los últimos autores, West redujo aún más el listado de voces. Finalmente, con la ayuda de James Endicott, West recogió un total de 1490 voces a partir de las definiciones de 24 000 lemas. Según el lexicógrafo, un estudiante de una lengua distinta al inglés o un aprendiente de inglés como lengua extranjera podría leer cualquier texto en dicho idioma sin la ayuda de un diccionario bilingüe, esto solo si poseyera el léxico registrado en su obra (Marco, 1997: 43).

No obstante, las obras de Ogden y Richards (1928) y West (1941) no contenían palabras de uso corriente en la comunicación natural del inglés, lo que representaba una debilidad, porque, si bien ambos trabajos cumplían con la meta de ser instrumentos adecuados para la enseñanza del inglés como lengua extranjera de forma sencilla y rápida, estos vocabularios distaban mucho de los usos cotidianos reales (Marco, 1997: 42). En el caso específico del texto de Ogden y Richards (1928), los hablantes promedio del inglés consideraban que estos léxicos se alejaban de la realidad lingüística, percibiéndolos como extraños respecto a la comunicación ordinaria (Sánchez-Saus, 2011: 51).

Aunado a lo anterior, los lemas que formaban parte de las obras de Ogden y Richards (1928) y West (1941) procedían –como se ha indicado– de textos académicos, lo cual ya configuraba un escollo. Pues, se ha demostrado que, elaborar un vocabulario reducido a partir de corpus compuestos

[1] Para más detalles, ver http://ogden.basic-english.org/basiceng.html

por materiales de nivel culto/formal, resultaba inconveniente, puesto que estos tendían a contener un léxico poco usual en las interacciones. Esto debido a que los diccionarios semasiológicos suelen excluir lexías por razones ideológicas, dialectales y estilísticas, muchas de las cuales son recurrentes en las conversaciones cotidianas. Además, las obras lexicográficas generales no se actualizan (mucho menos en la época anterior al internet) al mismo ritmo que el léxico.

En este punto, resulta relevante subrayar el papel que juegan los corpus en los estudios realizados desde algunas corrientes lingüísticas, como la funcional, o en los enmarcados en la lexicografía y la lexicología. En este sentido, debe destacarse que la construcción de los corpus debe ser meticulosa, sobre todo, a partir de los avances de la lingüística de corpus y el procesamiento electrónico de datos lingüísticos (Núñez, 2021), puesto que, de esta manera, se resguarda la fiabilidad de los datos analizados. En este sentido, siguiendo los planteamientos de la lingüística de corpus –entendida como "una metodología para la investigación de las lenguas y del lenguaje, la cual permite llevar a cabo investigaciones empíricas en contextos auténticos y que se constituye en torno a ciertos principios reguladores poderosos" (Parodi, 2008: 96) – estos deben responder a criterios básicos, como: modo de comunicación, régimen de recopilación, lengua, temporalidad y representatividad (McEnnery y Hardy, 2012). Estas cualidades se definen a continuación.

En primer lugar, el modo de comunicación (*mode of communication*, McEnnery y Hardy, 3-4) se refiere al tipo y formato de los materiales que conforman los corpus. En el caso de los estudios de disponibilidad léxica, se trata de textos escritos, mayormente de fuente primaria. En segundo lugar, el régimen de recopilación del corpus (*data collection regimes*) compete al requisito de variedad de los materiales. Es decir, los datos deben provenir de los diferentes tipos y géneros textuales, de manera que pueda reflejar lo mejor posible la realidad de la lengua en un punto temporal y geográfico específico (McEnery y Hardie, 2012: 8). En tercer lugar, la lengua, concierne justamente al código o sistema lingüístico de los textos. En cuarto lugar, la temporalidad atañe a la organización de los datos en función del tiempo en el que fueron elaborados, de manera que se presenten los materiales en corte sincrónico o diacrónico. Por último, la representatividad, este es el criterio más complejo de definir, puesto que él confluye propiedades cuantitativas y cualitativas de los grupos. En palabras de Rojo (2021: 1) "el concepto de representatividad es bastante complejo, de modo que tenemos que limitarnos aquí a la idea de que el análisis del conjunto de textos integrados en un corpus debe dar una visión adecuada de aquello que pretende representar".

Retomando el tema principal de esta sección, el segundo método, *estadístico*, es el prominente en los estudios de disponibilidad léxica. Sobre él, Santos Palmou (2017: 112) afirma que: "El uso de

métodos lexicométricos para obtener las palabras más usuales reduce la arbitrariedad en la selección del vocabulario y permite además establecer cortes según su índice de uso". En este sentido, se resalta la confiabilidad de las investigaciones lexicológicas de corte empírico-cuantitativo, ya que la selección de las palabras no recae en la intuición lingüística de los lexicógrafos, sino en las precisiones de los cálculos matemáticos realizado a los datos. Es decir, los especialistas emplean fórmulas estadísticas (complejas) que ofrecen valores –teóricamente ajustados a la realidad lingüística explorada– que permiten determinar la importancia de una palabra X sobre otra Y dentro del corpus. Así pues, las lexías de los diccionarios estadísticos se ajustan lo más imparcialmente posible a los usos reales de la lengua. A la luz de las observaciones previas, entonces, puede acotarse que, si bien las pesquisas de corte cuantitativo no están exentas de errores, estos suelen ser menores en comparación con los de trabajos llevados a cabo mediante otro marco analítico, como el lógico.

Una de las ventajas de desarrollar un trabajo desde la lexicografía estadística es contar con índices que contribuyen con la descripción de las propiedades del léxico de una sintopía. Respecto a los índices, Butrón (1991: 80) señala que estos "[…] son medios para ordenar las unidades léxicas mediante fórmulas matemáticas". En este sentido, los diccionarios estadísticos, como los de disponibilidad léxica, ordenan jerárquicamente los lemas a partir de los valores que muestra el índice utilizado. Particularmente, los lexemas –en términos de Lyon (1997) – no se organizan alfabéticamente, sino por el grado de relevancia indexado. Además, en contraste con los diccionarios semasiológicos, los estadísticos no presentan definiciones. En relación con lo anterior, Castillo Fadić (2021a: 19) aclara que este tipo de trabajo se denomina así "No porque tenga como tema la estadística ni porque esté dirigido a expertos matemáticos, sino porque se basa en cálculos estadísticos y presenta, en vez de las definiciones que solemos encontrar en la mayoría de los diccionarios, índices estadísticos". Con la finalidad de ilustrar la definición anterior, en la Figura 1, se muestra un fragmento de la primera página de *French Word List* de Henmon (1924), una de las primeras obras léxico-estadísticas. En esta página se aprecia la propiedad no definitoria de los diccionarios estadísticos, en los que, si bien no se incluyen los significados y acepciones de los lemas, sí se señalan los índices matemáticos de cada unidad lingüística. Además, cada palabra aparece ordenada jerárquicamente de mayor a menor en relación con su rango, mas no por el orden alfabético.

Figura 1. Primera página con las palabras más frecuentes del *French Word List*, parte I

FRENCH WORD LIST
PART I

le, la, l'	27,749	autre	695
de, du, des	21,948	petit	686
à, au, aux	8,581	vouloir	601
être (v.)	8,242	donner	577
et	7,628	savoir	561
un	7,381	quand	547
je, me, moi	6,021	celui, celle	547
avoir	5,488	votre	522
il, ils	5,459	venir	503
que (conj.)	5,129	bon	483
ce, cette, cet, ces	4,800	prendre	481

Fuente: Henmon (1924: 7)

Si bien se ha mencionado que la disponibilidad léxica se inserta en los estudios basados en el método estadístico, esta no es la única línea de esta área investigativa. En este marco también se incluyen las investigaciones sobre diccionarios de frecuencia y de léxico básico (Marco, 1997: 30). Con el fin de ahondar en las aplicaciones de la estadística en la lexicografía y contextualizar los estudios de DL en este panorama, en los siguientes epígrafes se reseñan someramente las características generales de los diccionarios de frecuencia léxica y léxico básico, los cuales se vinculan estrechamente con la línea de investigación de la DL.

1.1.1. Características generales de los diccionarios de léxico frecuente y léxico básico

Los diccionarios de frecuencias muestran los vocablos más usados de una comunidad de habla a razón del número de apariciones u ocurrencias que tengan en el corpus estudiado. Ezquerra (1974: 4) y Ávila Martín (2010: 164) apuntan a que los objetivos y aplicaciones de este tipo de trabajo han sido variados desde que empezaron a realizarse a finales del siglo XIX y principios del siglo XX. Sin embargo, han mantenido un enfoque mayormente pedagógico. Al respecto, Ávila Martín (2010: 163) afirma que: "Los estudios estadísticos del léxico se han aplicado a la enseñanza y aprendizaje de las lenguas desde comienzos del siglo XX". No obstante, desde una perspectiva metodológica pura (López Morales, 1994b), los diccionarios estadísticos también han permitido determinar las propiedades matemáticas del léxico de una lengua. Según Alvar Ezquerra (2004: 21), los objetivos de estos diccionarios se enumeran como sigue:

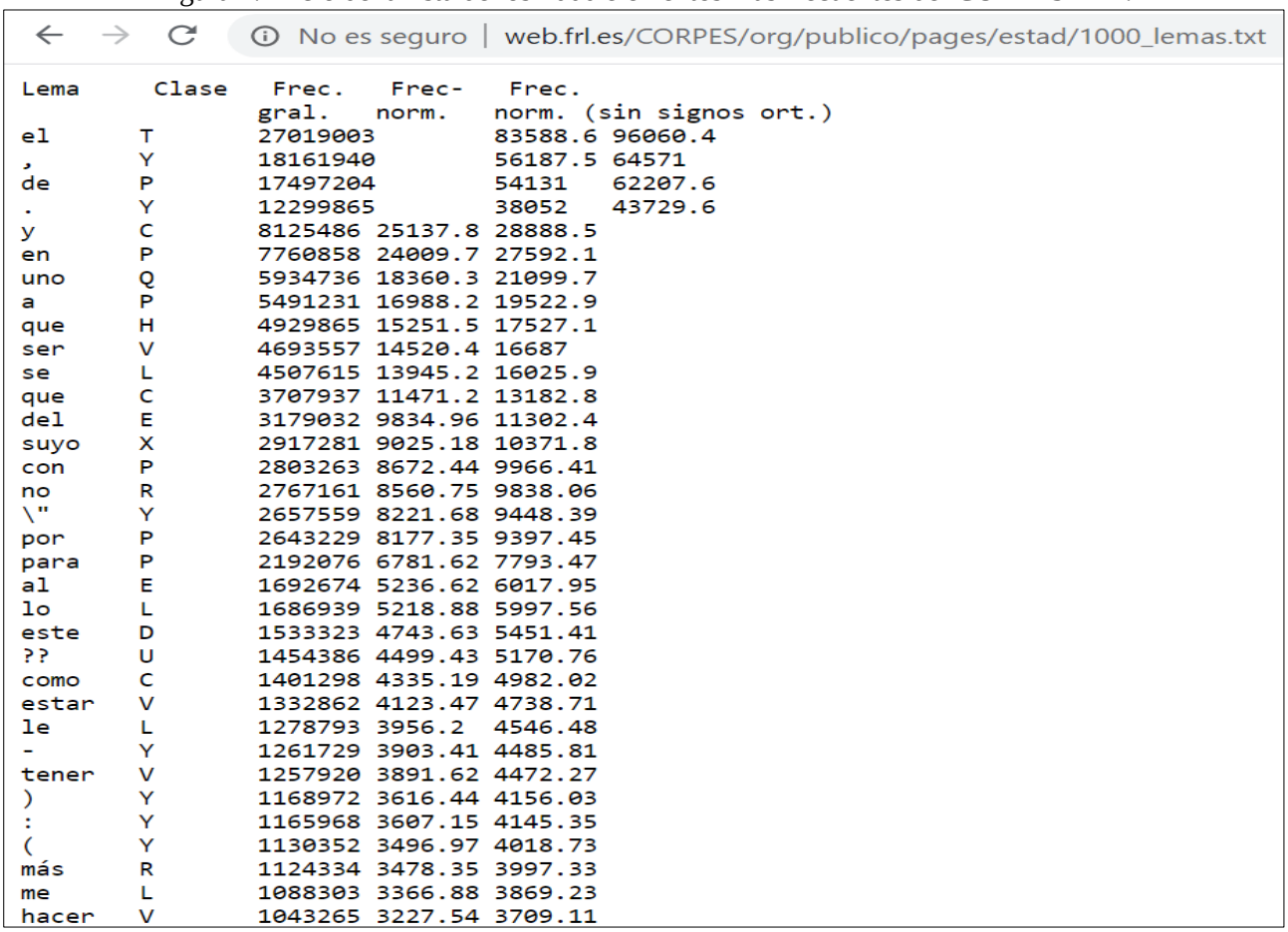
> 1° Conocer la importancia de las palabras de acuerdo con su frecuencia y distribución.
> 2° Decidir qué palabras deben enseñarse o no en cada nivel, en cuáles hay que insistir más y cuáles deben enseñarse para que los alumnos entiendan los materiales que manejan.
> 3° Poseer una información objetiva sobre las palabras y cuándo deben enseñarse.
> 4° Extraer el inventario de palabras del ámbito lingüístico en que se mueven los alumnos para facilitarles la comprensión de su entorno.
> 5° Saber cuáles son las palabras de ámbitos concretos cuando se enseña la lengua con fines específicos.
> 6° Seleccionar los grupos de palabras y formas que deben enseñarse conjuntamente, y en qué orden.
> 7° Determinar el grado de dificultad del léxico de los materiales de lectura y aprendizaje.
> 8° Que quienes confeccionan los materiales pedagógicos puedan seleccionar el léxico que debe aparecer en ellos.

Respecto a los criterios estadísticos, los investigadores han utilizado tres tipos de indicadores cuantitativos: 1) frecuencia, 2) rango y 3) mixto: combinación de los anteriores. No obstante, a la luz de las contribuciones de Juilland y Chang-Rodríguez (1964), la importancia de una palabra no radica únicamente en su frecuencia de aparición, sino también en su grado de dispersión. Puesto que esta mide la estabilidad de la frecuencia de un vocablo en el corpus; o, lo que es lo mismo, indica la distribución de una pieza léxica en los distintos tipos de textos bajo análisis. En este sentido, una lexía X con alta frecuencia es poco relevante, si solo aparece en uno o pocos textos. En cambio, una unidad léxica Y con alta ocurrencia, distribuida en un mayor número de textos, adquiere mayor importancia léxico-métrica (Ávila Martín, 2010: 169).

En el caso de los estudios de frecuencia léxica, hay que resaltar dos aspectos del corpus: tamaño y tecnología. El primero remite a la necesidad de contar con grandes cantidades de textos para el análisis, de manera que los resultados sean lo más cercanos posible a la realidad lingüística de una comunidad. El segundo, nuevas tecnologías, resalta la ayuda brindada por los procesadores computacionales, ya que con ellos se analizan grandes cantidades de datos en poco tiempo de manera expedita. Además, gracias a la ciencia de la computación se realizan búsquedas léxicas de forma casi inmediata. Igualmente, pueden desarrollarse estudios sobre las ocurrencias de las voces en un corpus particular (cf. Alvar Ezquerra, 2004; Núñez, 2021). A manera de ejemplo, puede observarse la manera cómo se exponen los elementos léxicos del *Corpus del Español del Siglo XXI* (CORPES XXI) de la RAE, en relación con las frecuencias de cada unidad analizada. Si bien el CORPES XXI no se construyó específicamente en el marco de los estudios de lexicometría, permite realizar consultas acerca de las ocurrencias de las más de “286 millones de formas” que lo componen, las cuales no son solamente palabras, sino también signos de puntuación y caracteres de edición, todo esto gracias a las ventajas que ofrece la lingüística computacional en el manejo y procesamiento de los materiales

textuales. En la Figura 2 se exhibe un fragmento de la lista de los 1000 elementos más frecuentes del CORPES XXI[2].

Figura 2. Inicio de la lista de los 1000 elementos más frecuentes del CORPES XXI.



Lema	Clase	Frec. gral.	Frec-norm.	Frec. norm. (sin signos ort.)
el	T	27019003	83588.6	96060.4
,	Y	18161940	56187.5	64571
de	P	17497204	54131	62207.6
.	Y	12299865	38052	43729.6
y	C	8125486	25137.8	28888.5
en	P	7760858	24009.7	27592.1
uno	Q	5934736	18360.3	21099.7
a	P	5491231	16988.2	19522.9
que	H	4929865	15251.5	17527.1
ser	V	4693557	14520.4	16687
se	L	4507615	13945.2	16025.9
que	C	3707937	11471.2	13182.8
del	E	3179032	9834.96	11302.4
suyo	X	2917281	9025.18	10371.8
con	P	2803263	8672.44	9966.41
no	R	2767161	8560.75	9838.06
\"	Y	2657559	8221.68	9448.39
por	P	2643229	8177.35	9397.45
para	P	2192076	6781.62	7793.47
al	E	1692674	5236.62	6017.95
lo	L	1686939	5218.88	5997.56
este	D	1533323	4743.63	5451.41
??	U	1454386	4499.43	5170.76
como	C	1401298	4335.19	4982.02
estar	V	1332862	4123.47	4738.71
le	L	1278793	3956.2	4546.48
-	Y	1261729	3903.41	4485.81
tener	V	1257920	3891.62	4472.27
)	Y	1168972	3616.44	4156.03
:	Y	1165968	3607.15	4145.35
(	Y	1130352	3496.97	4018.73
más	R	1124334	3478.35	3997.33
me	L	1088303	3366.88	3869.23
hacer	V	1043265	3227.54	3709.11

En cuanto a los estudios de léxico frecuente, hay que referirse al *Diccionario de frecuencias de las unidades lingüísticas del castellano*, de José Ramón Alameda y Fernando Cuetos (1995). El objetivo de los investigadores era determinar las ocurrencias y usos de las distintas unidades lingüísticas de la lengua castellana, por lo que no solo se centraron en el nivel léxico, sino también en los otros niveles de lengua (sintáctico, morfológico y fonético), además de los aspectos gráfico. Con esto, por un lado, los especialistas buscaban contribuir a la discriminación de variables en investigaciones de distintas disciplinas lingüísticas; y, por otro lado, sumaban esfuerzos en la enseñanza de lengua. En la Figura 3, se presenta un fragmento del diccionario de Alameda y Cuetos (1995: 118).

[2] Por razones de espacio no se exponen todos los 1000 elementos, pero pueden consultarse en https://www.rae.es/corpes/assets/rae/files/1000_lemas.txt

Figura 3. Fragmento del inicio del listado de palabras de la letra A de un diccionario de frecuencias.

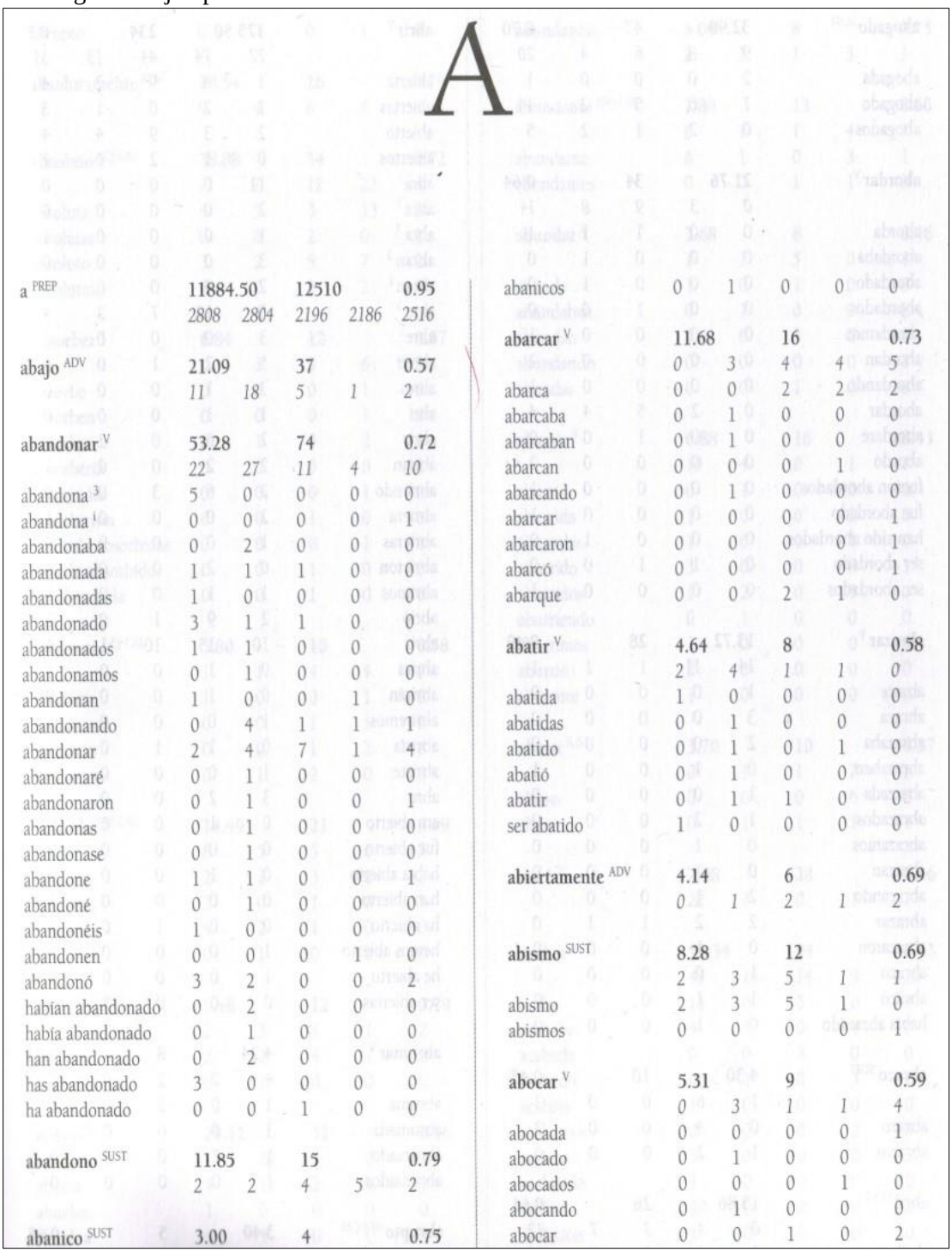
A

a:	44456	abandonase:	2	abastecido:	1
ababoles:	3	abandonasen:	1	abastecidos:	1
abacial:	1	abandonaste:	1	abastecimiento:	7
abad:	21	abandone:	9	abasto:	4
abades:	4	abandoné:	3	abastos:	2
abadesa:	3	abandonemos:	1	abate:	7
abadía:	3	abandonen:	1	abaten:	1
abadillo:	1	abandones:	3	abatí:	2
abajo:	304	abandono:	65	abatía:	2
abalanza:	3	abandonó:	45	abatían:	1
abalanzaban:	1	abandonos:	2	abatible:	1
abalanzado:	1	abanicaba:	2	abatida:	5
abalanzamos:	1	abanicando:	1	abatidas:	5
abalanzan:	1	abanicándose:	2	abatido:	11
abalanzándonos:	1	abanicaron:	1	abatieron:	1
abalanzaron:	2	abanicarse:	2	abatimiento:	6

Fuente: Alameda y Cuetos (1995: 118)

El segundo tipo de diccionario estadístico reseñado en estas líneas es el de léxico básico. Para Ávila Muñoz (1997: 25) el léxico básico se refiere a "los hábitos lingüísticos y de la competencia que los usuarios tengan sobre la lengua", el cual puede representarse como un entramado de un número limitado de unidades léxicas, generalmente "atemáticas", que se encuentran estadísticamente estables. En este sentido, el léxico básico recoge los vocablos más recurrentes en una comunidad, según un índice matemático que combina las ocurrencias y la dispersión de una expresión léxica (cf. Ávila Muñoz, 2006: 25). Así pues, los lemas que lo componen son los más frecuentes y, a su vez, presentan una mayor distribución en los diferentes tipos de textos que componen el corpus. En este tipo de trabajo, los materiales lingüísticos se organizan en relación con el género o mundo en el que se inscriben (López Chávez, 1994: 69). Los mundos conciernen a la tipología textual de los materiales del corpus, tales como: *periodístico*, *narrativa*, *técnica*, *ensayo* y *teatro* (González *et al.*, 1982). En la Figura 4 se ilustra un diccionario de léxico básico.

Figura 4. Ejemplo de un diccionario de léxico básico: inicio del listado de la letra A.



A

a PREP	11884.50		12510		0.95
	2808	*2804*	*2196*	*2186*	*2516*
abajo ADV	21.09		37		0.57
	11	*18*	*5*	*1*	*2*
abandonar V	53.28		74		0.72
	22	*27*	*11*	*4*	*10*
abandona 3	5	0	0	0	0
abandona 1	0	0	0	0	1
abandonaba	0	2	0	0	0
abandonada	1	1	1	0	0
abandonadas	1	0	0	0	0
abandonado	3	1	1	0	0
abandonados	1	1	0	0	0
abandonamos	0	1	0	0	0
abandonan	1	0	0	1	0
abandonando	0	4	1	1	1
abandonar	2	4	7	1	4
abandonaré	0	1	0	0	0
abandonaron	0	1	0	0	1
abandonas	0	1	0	0	0
abandonase	0	1	0	0	0
abandone	1	1	0	0	1
abandoné	0	1	0	0	0
abandonéis	1	0	0	0	0
abandonen	0	0	0	1	0
abandonó	3	2	0	0	2
habían abandonado	0	2	0	0	0
había abandonado	0	1	0	0	0
han abandonado	0	2	0	0	0
has abandonado	3	0	0	0	0
ha abandonado	0	0	1	0	0
abandono SUST	11.85		15		0.79
	2	2	4	5	2
abanico SUST	3.00		4		0.75
abanicos	0	1	0	0	0
abarcar V	11.68		16		0.73
	0	*3*	*4*	*4*	*5*
abarca	0	0	2	2	2
abarcaba	0	1	0	0	0
abarcaban	0	1	0	0	0
abarcan	0	0	0	1	0
abarcando	0	1	0	0	0
abarcar	0	0	0	0	1
abarcaron	0	0	0	0	1
abarcó	0	0	0	0	1
abarque	0	0	2	1	0
abatir V	4.64		8		0.58
	2	*4*	*1*	*1*	*0*
abatida	1	0	0	0	0
abatidas	0	1	0	0	0
abatido	0	1	0	1	0
abatió	0	1	0	0	0
abatir	0	1	1	0	0
ser abatido	1	0	0	0	0
abiertamente ADV	4.14		6		0.69
	0	*1*	*2*	*1*	*2*
abismo SUST	8.28		12		0.69
	2	*3*	*5*	*1*	*1*
abismo	2	3	5	1	0
abismos	0	0	0	0	1
abocar V	5.31		9		0.59
	0	*3*	*1*	*1*	*4*
abocada	0	0	0	0	1
abocado	0	1	0	0	0
abocados	0	0	0	1	0
abocando	0	1	0	0	0
abocar	0	0	1	0	2

Fuente: Castillo Fadić (2021a: 107)

Aunado a lo anterior, las investigaciones referentes al léxico básico han complementado los análisis con la aplicación del *índice de uso léxico*, que se calcula multiplicando la frecuencia por la dispersión (U = F x D). Este otorga un índice que refleja la importancia de cada palabra en el corpus,

con lo que, además, se determinan los vocablos que serán incluidos en el diccionario (Castillo Fadić, 2021a: 33). En esa misma línea, puede decirse que este tipo de vocabularios presenta un perfil más fiel y adecuado de las palabras usuales de una sintopía. En efecto, Santos Palmou (2017: 21) afirma que:

> De esta forma el léxico básico trabaja la frecuencia ponderada con la dispersión para ofrecer índices de usos; es decir, en lugar de juntar todo tipo de texto en un mismo grupo y sacar de ahí las frecuencias generales de las palabras que lo integran, divide todo ese universo léxico en diferentes mundos […] En consecuencia, sus resultados constituyen cálculos de las frecuencias de uso real de las distintas clases de palabras, dando preferencia a las más estables.

En los diccionarios de léxico básico, los vocablos con los mayores índices suelen ser palabras funcionales o gramaticales, como preposiciones, conjunciones, artículos, pronombres, ya que estas son más ocurrentes en los materiales que integran los corpus. Empero, también se encuentran lexías de otras categorías, como verbos, sustantivos y adjetivos. Sin embargo, sus índices generalmente son bajos, ya que no aparecen recurrentemente en los textos analizados (Ávila Martín, 2010: 172). En relación con lo mencionado previamente, Benítez, Hernández y Samper Padilla (1995: X) afirman que:

> el léxico básico, por el contrario, recoge las palabras más frecuentes, aquellas que aparecen continuamente en los discursos, con independencia del tema que se trate. Como es lógico, en los primeros lugares de los léxicos básicos aparecen siempre las palabras gramaticales; los sustantivos, dada su menor estabilidad estadística (como corresponde a su mayor concreción semántica), son las palabras menos frecuentes.

En el mundo hispánico, las investigaciones de disponibilidad léxica son abundantes, en comparación con las de léxicos básicos, según el arqueo bibliográfico. El trabajo icónico en este campo es el *Léxico Básico del español de Puerto Rico*, de Amparo Morales (1986), al que se le pueden sumar el: *Léxico básico de la lengua escrita en la República Dominicana*, de Antonio González, Santiago Cabanes y Francisco García, de 1982; *Léxico básico de los niños preescolares costarricenses*, de Marielos Murillo y Víctor Sánchez, del 2006. Y, más recientemente, el *Léxico Básico del español de Chile*, de Natalia Castillo Fadić (2021a).

El vocabulario disponible y el básico son complementarios, puesto que, en conjunto, reflejan la realidad léxica de una sintopía. Además, ambos constituyen el léxico fundamental de una comunidad de habla (Benítez, Hernández y Samper Padilla, 1995). Los diccionarios fundamentales se refieren al caudal léxico necesario que un individuo debe conocer y manejar para poder desenvolverse eficazmente en cualquier interacción cotidiana (López Chávez, 1999: 43).

Por último, este apartado ha permitido ubicar *grosso modo* el contexto en el que se inscribe esta tesis. En los siguientes epígrafes se definen los términos y conceptos fundamentales de la disponibilidad léxica, partiendo con la distinción entre esta disciplina y léxico disponible.

1.2. Léxico disponible *vs*. disponibilidad léxica

En el marco del área de la léxico-estadística, resulta pertinente tener en cuenta la distinción entre léxico disponible y disponibilidad léxica, como advierten Hernández Muñoz y Tomé (2017) y Gómez Molina (2021). Por una parte, el LD se refiere a las piezas léxicas que un hablante produce en un momento particular, ya que están potencialmente accesibles en el lexicón para ser utilizadas en la interacción, cuando se trata un tópico concreto en la interacción.

Por su parte, la DL corresponde a "una herramienta que tenemos a nuestra disposición para arrojar luz sobre el proceso de evocación de las palabras disponibles y entender cómo es el nivel lingüístico del hablante" (Gómez Molina, 2021: 207). Por consiguiente, la DL compete al aparataje teórico y metodológico por medio del cual se obtiene y explica el caudal léxico individual y colectivo (Hernández Muñoz y Tomé, 2017: 100). En relación con esta investigación, la diferenciación de los términos DL y LD resulta importante, porque permite delimitar, por un lado, los datos analizados y el contenido del diccionario estadístico; y, por otro lado, particularmente, los métodos y técnicas de recolección, organización y procesamiento de los datos.

En relación con lo anterior, esta tesis se sustenta en los postulados de la disponibilidad léxica. Del diseño metodológico, deben resaltarse los métodos de recolección de datos, ya que se han propuesto dos técnicas para la recopilación del léxico disponible de la población bajo análisis. El primero compete al método planteado por el PPHLD, que consiste en tomar los datos *in situ*, a través de cuestionarios de papel; en estos los participantes escriben a mano sus listas de palabras, para lo cual cuentan con dos minutos por cada centro de interés. Esta forma es una adecuación de la metodología utilizada por Gougenheim *et al.* (1964). En los párrafos siguientes se referirá a este método como *formato* o *técnica tradicional*.

El segundo método –denominado *formato* o *técnica tradicional*– ideado para esta tesis estriba en recoger los datos de manera remota o asincrónica, por medio de una página web *ad hoc*, en la que los encuestados elaboran sus listas de palabras y, al finalizar, quedan registradas en una plataforma electrónica, a la que tiene acceso únicamente el investigador responsable. Si bien se trata de un método distinto al formato tradicional, este ha intentado guardar, lo más fiel posible, las características y criterios de metodológicos esbozados por el PPHLD.

La disponibilidad léxica es un campo disciplinar que, cabalga entre la lingüística y la psicología, como afirma Hernández Muñoz (2006). Es eminentemente lingüística, debido a que su objeto de estudio es el léxico de la lengua, por lo que, además, se enraíza en la lexicología y la lexicografía. Como parte de las ciencias del lenguaje, busca conocer el vocabulario de una comunidad, para lo cual emplea una metodología con un componente altamente cuantitativo. En tanto que, del lado de la psicología, a través de la DL se pretende conocer la organización y activación de las palabras en el lexicón mental. Adicionalmente, se observa la adquisición y ampliación del caudal léxico. En efecto, de acuerdo con teorías provenientes de la psicología cognitiva, se construye el *léxico disponible colectivo*, que resulta de la "suma de un número significativo" de los *léxicos disponibles individuales* de una comunidad o sintopía (Hernández Muñoz y Tomé, 2017: 100). Esta sumatoria da cuerpo a los diccionarios de disponibilidad.

No obstante, debe declarase que, aunque navegue entre dos aguas, la DL se ancla en la lengua, por lo que el tema de esta investigación se orienta desde la lingüística, con una perspectiva sociolingüística. Así pues, en el siguiente subapartado, se definen los términos básicos sobre los que se sustenta esta disciplina, tales como: *palabra*, *vocablo* y *centro de interés*.

1.3. Conceptos fundamentales de la disponibilidad léxica

1.3.1. Lexicografía y lexicología

La inquietud del hombre por atestiguar y ordenar el léxico de su propia lengua, por un lado, contrastar su vocabulario con el de sistemas lingüísticos distintos al suyo, por otro lado, impulsó el nacimiento de dos disciplinas lingüísticas que –aunque están hermanadas– poseen características distintivas, a saber: lexicografía y lexicología. Ambas comparten el mismo objeto de estudio: el léxico. No obstante, el acercamiento a este se lleva a cabo desde ángulos separados. La primera da cuenta de él a través de descripciones; en cambio, la segunda "se preocupa por los principios y leyes generales que rigen el vocabulario" (Porto Dapena, 2002: 17). Entonces, la lexicografía tiene el objetivo principal de elaborar diccionarios mediante técnicas científicas. Hay que acotar que este carácter científico es un atributo recientemente adquirido, porque desde sus orígenes –y hasta bastante entrada la era contemporánea– se catalogaba más como un arte u oficio, y no como una práctica ligada a la ciencia (Pérez, 2005). En contraparte, la lexicología posee una naturaleza más teorética. A pesar de las discusiones acerca del carácter más/menos científico de la lexicografía *versus* la lexicología –como reseña Porto Dapena (2002)–, debe acentuarse la relación complementaria de las dos en el abordaje del vocabulario de una lengua.

1.3.2. Palabra y vocablo

El núcleo de la lexicología y la lexicografía se encuentra precisamente en la *palabra*, la cual "desempeña un papel tan crucial en la estructura de la lengua que precisamos de una rama especial de la lingüística para examinarla en todos sus aspectos (el origen, la forma y el significado). Esta rama se llama Lexicología" (Ullmann, 1962: 33, en Cardero, 2011: 7). La palabra como unidad de análisis lexicológica no está exenta de problemas teóricos. A la inversa, parafraseando a de Miguel (2009: 13-14) y a (Piera, 2009: 25), por ser un objeto de estudios de interés de otras disciplinas lingüística (morfología, semántica, sintaxis) y, además, abordado desde distintas perspectivas (psicolingüística, neurolingüística, sociolingüística, lingüística de corpus, entre otras), definirla y caracterizarla resulta avasallante.

Al respecto, Felíu (2009: 54-55) afirma que "[…] Resulta difícil encontrar un único criterio de definición de la unidad 'palabra' que resulte válido tanto desde un punto de vista interlingüístico como desde la perspectiva de una lengua concreta". Es tanta la importancia de las palabras –como elementos de la lengua– en la vida de los seres humanos que está cargada, no solo de significados y conceptos lingüísticos, sino también de significados culturales y experienciales. Además, dentro de lenguas, como el español, su función sobrepasa los niveles de análisis, puesto que puede ser analizada de forma diferente en cada uno de ellos. Por ejemplo, es la unidad de análisis mayor de la morfología; y es también uno de los componentes cardinales de la sintaxis.

En la Nueva gramática de la lengua española, se indica que: "La PALABRA constituye la unidad máxima de la morfología y la unidad mínima de la sintaxis. El concepto de 'palabra' está habitualmente ligado a la representación gráfica de la lengua, ya que las palabras van separadas por blancos en la escritura" (RAE, 2009: 11). La Academia manifiesta en dichas líneas uno de los principios esgrimidos en algunos círculos lingüísticos, según los cuales se determina la palabra a partir de las separaciones virtuales habidas entre uno y otro signo lingüístico. Sedano (2011: 50) concuerda con esta perspectiva al precisar que la palabra "Es una unidad entre pausas virtuales. Estas pausas permiten aislar la palabra, al menos metodológicamente, del resto de los constituyentes sintagmáticos." Pero, la lingüista, considera inmediatamente que –visto de esta manera– el término funcionaría, adecuada y únicamente, en los análisis gramaticales, sobre todo aquellos basados en corpus (escritos). Empero esta óptica cabría aplicarla solamente en las lenguas con escritura. Además, esta concepción del término palabra enfrentaría algunos tropiezos analíticos, como, por ejemplo, los clíticos en español.

Por su parte, Mendívil Giró (2009) distingue entre palabras "simples" y *palabras con estructura externa* (PPEE). Según el autor, una PPEE "es un sintagma que funciona a ciertos efectos semánticos y formales como una palabra única" (Mendívil Giró, 2009: 84). A grandes rasgos, pareciera que este término enfocaría el análisis de las piezas léxicas desde una perspectiva fraseológica. Sin embargo, a tenor de la argumentación del lingüista, se trataría de un terreno más amplio en el *continuum* de los diversos tipos de expresiones complejas que poseen las lenguas (Mendívil Giró, 2009).

En este punto resulta apropiado enfocar la mirada hacia las construcciones léxicas complejas; es decir, aquellos elementos constituidos por dos o más palabras. El abordaje de este tipo de unidades léxicas puede llegar a ser uno de los terrenos más espinoso de la lingüística, particularmente de la gramática, lexicografía, terminología, fraseología y, por supuesto, de la disponibilidad léxica. En efecto, García-Page (2008: 38) arguye que:

> […] uno de los más espinosos problemas de la delimitación entre locuciones nominales y colocaciones (y compuestos) se plantea en la observación real de ciertas combinaciones de difícil clasificación: lo que para unos es locución, para otros es un compuesto o colocación.

Desde un primer plano, las piezas léxicas complejas pueden caracterizarse desde la composición, como mecanismo de formación de palabras (cf. Aitchison, 1993; Alvar, 1996; Di Tulio 1997; Almela Pérez, 1999; Otaola, 2004; RAE-ASALE, 2009; Pavey, 2010; Sedano 2011; Del Barrio y San Vicente 2015). Este procedimiento consiste en formar una nueva palabra a través de la unión de dos o más lexemas independientes, con un significado unitario (Lara, 2006; Sedano, 2011). En algunas taxonomías se diferencian los compuestos *propios* e *impropios*. Los primeros atañen a las piezas léxicas unidas sin ningún tipo de separación o interrupción formal y, a su vez, presentan una pronunciación unitaria, por ejemplo: *aguamarina* y *abrelatas*. Los segundos, compuestos impropios o sintagmático, competen a los lexemas que no están amalgamados ni tienen una pronunciación unitaria, por ejemplo: *palo de agua* y *ojo de buey* (Sedano, 2011). Los compuestos sintagmáticos se distinguen de los sintagmas libres en que estos últimos pueden ser modificado –según los mecanismos de la sintaxis española–, ampliando su significado (ej. *La casa de María*, *la casa verde de María*, *la gran casa verde de María*); opuestamente, las palabras compuestas no pueden alterarse; así, *ojo de buey*, sincrónicamente, no puede cambiarse en *#gran ojo de buey* u *#ojo negro de buey* (cf. Sedano 2011). Sin embargo, respecto a la composición impropia, Lara (2006) las denomina sintagmas fijos, indicando que remiten a "[…] unidades de denominación compuestas por dos o más núcleos morfemáticos cuyo significado resulte de la composición de los significados de los núcleos-palabras tomados aisladamente", por ejemplos: *máquina de coser*.

En cambio, desde un segundo plano, las construcciones complejas son vistas como colocaciones, sobre todo desde la terminología teórica. Parafraseando a Ramírez (2019: 146), estas se refieren a las unidades léxicas integradas por dos o más palabras. En este orden de ideas, la combinación de los elementos léxicos en una colocación se basa en la frecuencia y en la preferencia; además, no actúan en bloque, sino que mantienen una relación sintagmática entre ellos. Sin embargo, delimitar y definir una colocación tampoco ha sido una tarea sencilla. En efecto, Bosque (2001) muestra un conjunto de marcos y aproximaciones conceptuales desde los cuales se ha intentado responder a la pregunta ¿qué es una colocación?, como: la inestabilidad, la unidad semiidiomática, la estadística y la preferencia, los cuales provienen, en gran medida, de la fraseología. No obstante, el lingüista insiste en que las colocaciones deben estudiarse desde el ámbito de la interfaz léxico-sintaxis y, más concretamente, desde las propuestas de selección léxica. Esta última estima que los predicados "[…] seleccionan a sus argumentos, y al hacerlo restringen el conjunto de entidades que pueden denotar en función de rasgos semánticos que vienen a ser muy abiertos unas veces y considerablemente restringidos otras" (Bosque, 2001: 2).

En consideración a los puntos expuestos someramente en los acápites previos, puede señalarse –sin pretensiones teóricas y conceptuales, sino más bien en función de la metodología de esta tesis doctoral– que las colocaciones son estructuras o elementos plurilexicales, constituidas por dos o más palabras. En estas se identifican –no siempre de manera sencilla– una base o núcleo y el colocativo o complemento (Bosque, 2001; Muñoz Núñez, 2010).

En los estudios de disponibilidad léxica no es imprevisto la aparición de elementos plurilexicales en los listados de palabras, al contrario, como menciona Ramírez (2019: 146) "[…] las respuestas que pueden arrojar los test de disponibilidad léxica abarcan tanto las construcciones simples (una sola palabra por unidad léxica) como las complejas (más de una palabra por unidad léxica)". Efectivamente, en el corpus de estudio se hallaron formas como: *mundo fantástico*, *establecimiento educacional*, *plan de estudio*, *matar el tiempo*, entre otros.

En atención al tema de esta tesis, el término palabra es tratado en consonancia con los postulados de la estadística léxica, con especial énfasis de la DL. Entonces, se parte de la distinción entre palabra y vocablo. Así pues, el término *palabra* se entiende como la unidad mínima del análisis textual, mientras que el elemento vocablo se refiere a la unidad mínima del léxico (Atkins y Rundell, 2008; Castillo Fadić y Sologuren, 2020). Aún más explícitamente, la palabra se enmarca en un contexto de producción lingüística, por lo que puede juzgarse como un elemento –por así decirlo– real, que se cristaliza a través de los sonidos y los grafemas de la lengua. En cambio, el *vocablo* se

ubica en el plano abstracto de la lengua, al alcance de todos los hablantes. En este sentido, el término posee un carácter más teórico (Castillo Fadić, 2021a).

A manera de ejemplo, cuando una persona lee las noticias en la prensa, se topa visualmente con *palabras*; es decir, elementos léxicos concretos, materializados en los textos (considérese la expresión *satisfizo*). Por el contrario, cuando el mismo sujeto consulta un diccionario con el fin de solventar dudas sobre la expresión antes mencionada, se encuentra con *vocablos*; a saber, las maneras como la lengua registra los signos (en este caso, la forma *satisfacer*). Así pues, un vocablo es la unidad lingüística que agrupa todas las variantes de una raíz derivada y/o flexionada (Lara, 2006; Ávila Muñoz y Lasarte, 2010). Entonces, en la lista hipotética (1), a continuación, se observan las siguientes expresiones:

(1) *casa*, *casas*, *casita*, *casitas*, *casucha*, *casota*.

Sobre la base de la discusión previa, puede indicarse que en (1) se contabilizan 6 palabras –o en términos de Lyon (1997) *expresiones de palabras*–. Sin embargo, tan solo se ha enumerado un único vocablo –*forma de palabra* (Lyon, 1997)–. Esto debido a que los seis elementos lingüísticos comparten la misma raíz, diferenciándose por medio de los afijos.

En relación con la metodología de la disponibilidad léxica, el término *palabra* refiere, entonces, al número de actualizaciones concretas que realizan los participantes en los listados. Por su lado, *vocablo* remite a las lexías producidas por los encuestados que, indistintamente del número de repeticiones que se observen en el corpus, cuentan como una sola unidad léxica (Gallego, 2014; Zhou, 2021). Al respecto, debe aclararse que el término vocablo, aquí asumido, engloba también a las unidades plurilexicales, como *mesa de profesor*, que se distingue de *mesa* a secas. En este sentido, en un listado hipotético como (2), el análisis cuantitativo de número de palabras y vocablos contabilizará 6 palabras y 4 vocablos.

(2) *silla*, *mesa*, *mesa de profesor*, *lápiz*, *silla*, *mesa*

Sin embargo, para autoras como Martínez (2007: 148), los términos palabra y vocablo pueden conllevar a confusiones entre lectores poco especializados, ya que tienen un carácter aparentemente sinonímico, debido a la definición que la RAE (2022, versión electrónica) entrega sobre el lema vocablo = palabra. En este contexto, la lexicógrafa –desde una perspectiva ecléctica fundada en la lingüística de corpus– propone utilizar los términos *tokens* (total de ocurrencias en los corpus) y *types* (tipos léxicos diferentes), con base en la operatividad de los datos, como etiquetas subsidiarias o complementarias de palabra y vocablo.

1.3.3. Palabras frecuentes, comunes y usuales

En la base epistémica de la disponibilidad léxica se encuentran los conceptos *palabras frecuentes*, *comunes* y *usuales*. A partir de la discriminación terminológica entre ellas, desde el enfoque léxico-métrico, pudo conocerse y entenderse el concepto de disponibilidad. Las características generales de estos tres tipos de palabras se relacionan con: i) la estabilidad estadística que posean en los corpus; ii) la función que cumplen en la gramática de la lengua, iii) el contenido semántico concreto que tengan, y iv) la utilidad que le dan los hablantes.

Gougenheim *et al.* (1954, 1964) aplicaron el método basado en la frecuencia (f) para determinar el vocabulario del corpus utilizado, debido a que era el procedimiento válido y, por ende, extendido en los trabajos léxico-estadísticos de esa época. En este contexto, se seleccionaban las unidades léxicas que aparecían más veces en los corpus. No obstante, los primeros recuentos del proyecto ofrecieron una realidad parcial y alejada a los objetivos de la empresa. Esto debido a que, en las primeras posiciones de los listados –organizados jerárquicamente de forma decreciente– aparecieron palabras gramaticales o funcionales, verbos y sustantivos de carácter general, y adjetivos. En cambio, los sustantivos concretos y comunes exhibían poca ocurrencia, por lo que quedaban excluido o, simplemente, no aparecían. A manera de ilustración, el lema *jupe* 'falda' alcanzó una f = 9, en 5 textos; muy similar ocurrió con *auto*, *métro*, *coude*, *dents*, *veston*, *fourchette*. Por su parte, la palabra *avoir* tenía una alta estabilidad, con un f = 11 552, repartido en 163 textos Gougenheim *et al.* (1964).

Lo anterior demostró que los términos *palabra frecuente*, *usual* y *común* no eran sinónimos. Cuantitativamente, las primeras conciernen a las que exhiben una alta ocurrencia en los corpus, mientras que las comunes y usuales se refieren a las unidades léxicas cuyos significados son conocidos y compartidos por una sintopía y, por ende, suelen ser utilizadas cotidianamente cuando la ocasión lo requiere. Al respecto, Mateo García (1997: 29) afirma que "es evidente que mientras algunas palabras aparecen habitualmente en cualquier conversación o escrito, con independencia de su tema, otras están muy relacionadas con lo que, en terminología de Halliday (1973), llamaríamos el "campo de la interacción comunicativa"'. Entonces, una palabra como *pizarra*, que es común y usual en el contexto educativo, puede no ser frecuente en un corpus.

1.3.4. Palabras atemáticas y temáticas

En atención a las razones esgrimidas previamente, René Michéa (1950) postuló que el léxico está constituido por dos tipos de palabras, a saber: las *atemáticas* y las *temáticas*. Las primeras consisten en las piezas léxicas que aparecen con alta frecuencia en los corpus, independientemente, de los contenidos o tópico de los materiales, por lo que son estadísticamente estables. Por el contrario,

las segundas, *palabras temáticas*, son poco frecuentes, debido a que su aparición está condicionada por los temas tratados en los textos (escritos u orales). En este grupo se localizan, mayormente, los sustantivos de contenidos semánticos concretos y específicos, a la vez que verbos.

A partir de esta oposición entre los términos *atemático* y *temático*, Michéa definió las «palabras frecuentes» y las «palabras disponibles». El término *disponible* se argumentó sobre la premisa de que, aunque la expresión de una pieza léxica de contenido semántico específico no sea frecuente, son potencialmente disponibles para ser utilizada en la interacción cuando el tema así lo requiera (Gougenheim *et al.*, 1964: 145).

1.3.5. Pruebas de fluencia semántica

En este punto, fueron decisivos los experimentos realizados por Michéa. En primer lugar, tomando en cuenta la hipótesis de que el interés y la atención permiten la memorización de términos, el lexicógrafo aplicó una prueba a 10 estudiantes, quienes cursaban alemán como lengua extranjera. El ejercicio consistió en leerles un fragmento (de aproximadamente 800 palabras) de la obra "Tonnelier de Nuremberg"; luego les pidió que escribieran las 20 palabras que más les llamaron la atención. Los resultaron arrojaron –a diferencia de las listas de frecuencias– un número alto de sustantivos concretos y comunes. En la figura 6, a continuación, se presentan las veinte palabras más recurrentes o disponibles de este experimento.

En un segundo experimento, Michéa les solicitó a sus estudiantes que escribieran las veinte primeras palabras que se les vinieran a la cabeza sobre el tópico "viaje en tren"[3]. Para esta tarea no había restricciones en el tipo de palabras a reportar. Los resultados mostraron, mayormente, sustantivos ligados al tema propuesto, por ejemplo: *gare*, *train*, *contrôleur* y *wagon*. En virtud de esto, Michéa indica que en este tipo de tarea está involucrada la memoria; por lo que la selección de una palabra potencial no está condicionada a una simple ley probabilística o al azar, sino a las relaciones de ideas y conceptos que una persona maneje sobre un tópico determinado. Así pues, las palabras disponibles están asociadas en la memoria semántica. Concretamente, Michéa (1953) afirma:

> En présence d'une *situation donnée*, les mots qui viennent les premiers à l'esprit sont ceux qui sont liés tout spécialement à cette situation et la caractérisent, c'est-à-dire les noms [...] La mémorie, basée sur l'association des idées, est sélective: elle n'obéit pas à une loi probabilitaire simple.

[3] Concretamente, la consigna utilizada fue: "Pensez à un voyage en chemin de fer à partir du moment où vous vous trouvez à la gare et écrivez les 20 mots qui vous viennent les premiers à l'esprit" (Gougenheim *et al.*, 1964: 147)

En atención a estos resultados, los investigadores concluyeron que las palabras con alto contenido semántico, como los sustantivos concretos, son estadísticamente estables en estos ejercicios. Textualmente, señalaron: “On peut constater que dans cette sorte d'enquête les mots concrets essentiels apparaissent avec une stabilité remarquable” (Gougenheim *et al.*, 1964: 148). En consecuencia, establecieron este tipo de pruebas como el instrumento adecuado para la recolección de los materiales del proyecto del *Francés fundamental*.

Figura 5. Las veinte palabras más disponibles del experimento de Michéa.

Voici la liste :

1	or	9	11	Paumgartner	5
2	ventre	9	12	solidité	4
3	Nuremberg	8	13	tambouriner	4
4	Martin	7	14	armoire	4
5	gros	7	15	foudre	4
6	tonnelier	6	16	grand	4
7	président	6	17	lentement	3
8	argent	6	18	lourdement	3
9	tonneau	6	19	assemblée	3
10	maître	5	20	richesse	3

Fuente: Gougenheim *et al.* (1964: 146)

1.3.6. Centros de interés

Conforme a lo anterior, Gougenheim et al. (1964) se concentraron en determinar el grado de disponibilidad que tenía una palabra X en un dominio conceptual Y. Por lo tanto, el enfoque no apuntaba a conocer, por ejemplo, si *cabeza* era más frecuente que *rosa*, sino en identificar, cuantitativamente, si *cabeza* era más disponible que *ojo*, *brazo*, *esófago* o *hígado* en el campo semántico *cuerpo humano*, a manera de ilustración (cf. Gougenheim *et al.*, 1964: 152). En este panorama, las reflexiones provenientes de la psicología experimental y la pedagogía acerca de cómo los grupos humanos comparten concepciones generales sobre tópicos particulares, jugaron un papel fundamental en el diseño de los instrumentos de recolección de datos (Ávila Muñoz y Sánchez, 2011: 47).

En el caso del área de la pedagogía, se utilizaban los análisis de palabras y conceptos para identificar las necesidades primordiales del alumnado, con el fin de adecuar los planes curriculares de los institutos. Para esto los docentes recurrían a pruebas de fluencia semántica, encabezadas por una consigna que activaba las ideas que los sujetos tenían y relacionaban con una categoría específica,

denominada *centro de interés*[4] (Sánchez Iniesta, 1995). Este último se entiende como los estímulos verbales que activan cognitivamente las palabras en la mente de los encuestados (Lara, 2006; Sánchez-Saus, 2011, 2019; Paredes García, 2014).

Entonces, el instrumento que sirvió para recoger los datos lingüísticos del proyecto francés es una prueba asociativa de alto contenido instruccional que enciende el léxico potencial de un hablante por medio de núcleos temáticos o centros de interés. A partir de estas disquisiciones, Gougenheim y el resto del equipo definieron las dieciséis áreas nocionales que debían ser analizadas; estas son[5]:

1) Las partes del cuerpo
2) La ropa (sin importar que sea ropa de hombre o de mujer)
3) La casa (pero sin los muebles)
4) Los muebles de la casa
5) Los alimentos y bebidas de las comidas (en todas las comidas del día)
6) Los objetos colocados sobre la mesa y de los que nos servimos en todas las comidas del día
7) La cocina, sus muebles y los utensilios que se encuentran en ella
8) La escuela, sus muebles y su material escolar
9) La calefacción y la iluminación
10) La ciudad
11) El campo y el burgo
12) Los medios de transporte
13) Los trabajos del campo y del jardín
14) Los animales
15) Los juegos y distracciones
16) Los oficios (los diferentes oficios y no los nombres que se refieren a un solo oficio)

En síntesis, la disponibilidad es la herramienta por medio de la que se llega al léxico disponible de una sintopía, por lo que –al centrar la atención en el léxico– se considera inscrita en la lexicología y lexicografía, especialmente en el campo de los estudios léxico-estadísticos. Sus unidades de análisis

[4] También suelen denominarse como: *áreas nocionales*, *ejes temáticos*, *estímulos verbales* y/o *actualizadores*.

[5] Nuestra traducción del francés: 1) Les parties du corps; 2) les vêtements (peu importe que ce soient des vêtements d'homme ou de femme); 3) la maison (mais pas les meubles); 4) les meubles de la maison; 5) les aliments et boissons des repas (à tous les repas de la journée); 6) les objets placés sur la table et que nous utilisons à chaque repas de la journée; 7) la cuisine, ses meubles et les ustensiles qui s'y trouvent; 8) l'école, ses meubles et son matériel scolaire; 9) le chauffage et l'éclairage; 10) la ville; 11) le village ou le bourg; 12) les moyens de transport; 13) les travaux des champs et du jardinage; 14) les animaux; 15) les jeux et distractions; y 16) les métiers (les différents métiers et non pas les noms qui se rapportent à un seul métier).

son la palabra y el vocablo. La primera, se entiende como las expresiones lingüísticas concretas producidas por los hablantes, mientras que la segunda, se refiere a la forma que aglutina todas las expresiones que comparten una misma raíz léxica (Lara, 2006). Estas unidades se recolectan a través de pruebas asociativas de alto contenido instruccional, integradas por estímulos verbales, denominados centros de interés (Sánchez-Saus, 2019).

Si bien, a través de todo el texto se ha mencionado el proyecto del Francés elemental como punto de partida de las investigaciones de disponibilidad, en los siguientes apartados se profundiza acerca de las características generales y específicas de la empresa llevada a cabo por George Gougenheim, René Michéa, Paul Rivenc y Aurélien Sauvageot.

1.4. El proyecto del Francés fundamental

El objetivo de este apartado es reseñar el contexto en el que se desarrollaron los trabajos del Francés elemental, como estudio fundacional de la disponibilidad léxica. Concretamente, 1) se dibuja el nicho del proyecto como parte de los planes educativos de la UNESCO; 2) se describe la metodología empleada; 3) se exponen los resultados más relevantes; y 4) se compara el proyecto con el inglés básico. Para cumplir estos objetivos, se ha realizado un arqueo bibliográfico, en el que se han resaltado las siguientes fuentes: Gougenheim (1954, 1955); Gougenheim *et al.* (1964); Marco (1997); Ávila Muñoz y Sánchez (2010); Sánchez-Saus (2011, 2019); Santos Díaz (2015, 2020); Callealta y Gallegos (2016); Santos Palmou (2017), entre otros.

Los estudios de disponibilidad léxica se iniciaron en Francia a mediados de la segunda mitad de siglo XX, en los trabajos preparatorios de un proyecto de lengua base o un francés básico, cuyos propósitos eran facilitar la enseñanza del francés y acelerar su difusión en el mundo. Por lo anterior, el foco estaba dirigido hacia tres grupos particulares (Gougenheim, 1955: 404-405):

a. *Los extranjeros*. Francia siempre había sido un destino turístico por excelencia, por lo que muchas personas alrededor del orbe tenían interés por visitarla y conocerla. Aunado a esto, muchos querían vivir en este país y empezar nuevos planes en él. En este contexto, era imperativo expandir el idioma de Juana de Arcos y Baudelaire y, paralelamente, recobrar la importancia que el francés había tenido en tiempos pretéritos.
b. *Los pobladores de la Unión Francesa*. A mediados del siglo XX, Francia aún contaba con colonias en todo el mundo. Sin embargo, en cada una de ellas se hablaban lenguas locales. En África, por ejemplo, las distintas comunidades que componían la Unión tenían múltiples lenguas y dialectos. En consecuencia, se hacía necesario contar con una lengua común que permitiera una mejor comunicación y relación entre los distintos pueblos de la Unión. Empero,

enseñar un francés "completo" sería una tarea titánica –especialmente para los adultos–, por lo que era ineludible contar con una gramática simplificada.

c. *Las personas que vivían en Francia, pero no conocían o no manejaban bien la lengua*. Al país europeo llegaban a trabajar personas de distintas nacionalidades, especialmente norafricanos, quienes solo hablaban su lengua materna o tenían un dominio muy bajo del francés. Y el poco tiempo libre del que disponían no les permitía asistir a clases de francés de larga duración, por lo que era perentorio disponer de materiales didácticos basados en una lengua base o reducida.

No obstante, estos primeros propósitos y destinatarios cambiaron desde la publicación de la primera edición del Francés elemental, en 1954, a la nueva y definitiva versión de 1964, que –parafraseando a los autores– presentaba estos nuevos objetivos:

a. La enseñanza escolar: práctica y cultura.
b. La difusión del francés como lengua nacional o lengua extranjera privilegiada, en los jóvenes estados independientes.
c. Cubrir las necesidades comunicativas de los viajeros, turistas, de los extranjeros que se movilizan en territorios de habla gala por distintas razones (educativas, comerciales, empresariales)

(Gougenheim *et al.*, 1964: 20-21)

Una de las primeras críticas que debió sortear el proyecto fue respecto al nombre inicial: *Francés Básico*. Este tuvo que descartarse a la brevedad, porque daba la impresión de que podría estar vinculado, teórica y metodológicamente, con el Inglés Básico, trabajo que no había sido bien recibido entre los especialistas (Gougenheim, 1954 y 1955). Así pues, tuvo que intitularse nuevamente, esta vez como: Le français élémentaire, nombre que se cambió a Français fundamental, en la reedición de 1959. Finalmente, en 1964, se publicó una nueva edición, ampliada y mejorada, con el título de 1959. Esta última versión estuvo compuesta por una lista de 1475 palabras, de las cuales 1212 eran de contenido léxico concreto; 253, gramaticales, y 10 interjecciones (Gougenheim *et al.*, 1964: 12-13).

Tomando en cuenta los antecedentes sobre elaboración de recuentos léxicos y vocabularios simplificados para la enseñanza de lengua, los investigadores franceses plantearon una metodología basada en tres principios (Gougenheim, 1955: 410; Gougenheim *et al.*, 1964: 13):

a. La frecuencia de palabras en la lengua hablada[6]
b. La búsqueda de las palabras disponibles más útiles
c. El empirismo racional

[6] Nuestra traducción de: 1) *la fréquence des mots dans la langue parlée*, 2) *la recherche des mots disponibles les plus utiles*, y 3) *l'empirisme rationnel*.

Sobre la primera dimensión, en la mayoría de los trabajos léxico estadísticos previos al Francés elemental, como se ha revisado desde la Introducción, los recuentos léxicos se basaban en el análisis de frecuencias de palabras de textos escritos. En algunos de ellos se tomaban en cuenta cartas personales, como ejemplos de materiales más cercanos a la oralidad y la lengua cotidiana. Pero este tipo de textos era poco representativo en los corpus de estudios en los que llegaba a utilizarse. En dirección opuesta a esta tendencia, los lexicógrafos consideraron óptimo analizar muestras orales. De esta manera, además de ayudar a la elaboración de materiales didácticos vanguardistas, favorecieron los avances metodológicos relacionados con la confección de los corpus modernos. Al incluir materiales orales, se buscaba que los listados mostraran un léxico más cercano al utilizado en la comunicación natural diaria. La razón se fundamentaba en el supuesto de que los textos orales se alejan de la sofisticación léxica a la que pueden llegar los autores de obras literarias y religiosas, editores, textos técnicos y científicos, entre otros.

El segundo rasgo constituye el centro de los estudios de disponibilidad léxica y representa no solo una contribución metodológica, sino también teórica sobre la caracterización del léxico. Esto, debido a que –en la preparación y posterior curso del Proyecto– los investigadores innovaron conceptos lexicológicos, a partir de los cuales formularon un método capaz de, primero, recoger las palabras más disponibles del francés y, segundo, seleccionar estadísticamente las palabras que conformarían el vocabulario reducido. Así pues, Gougenheim y su equipo contribuyeron no solo con una obra en el área de los vocabularios reducidos con fines didácticos, sino que, además, avanzaron en las propuestas metodológicas básicas que guiarían los subsiguientes trabajos sobre disponibilidad léxica y enseñanza de lengua.

1.4.1. Análisis estadísticos del francés fundamental

El proyecto del Francés elemental estuvo concebido como un trabajo lingüístico enmarcado en la léxico-estadística (Mateo García, 1997: 28). En esta línea de investigación –que ya contaba con una larga tradición, como se ha expuesto en los apartados previos– se asumía el índice de frecuencia como el criterio válido y más extendido para la determinación y selección de los vocabularios reducidos. Al respecto, Gougenheim (1954: 218) señala que “Notre première tâche a été de dresser une liste des fréquences. Les mots les plus nécessaires sont a priori ceux qui sont employés le plus fréquemment”.

Para llevar a cabo la elaboración del Francés elemental, los lexicógrafos franceses –a diferencia de sus antecesores– basaron sus análisis en un corpus de materiales orales. En los años 50

del siglo XX, ya se contaba con nuevas tecnologías, ausentes antaño, para la recopilación de materiales orales. De hecho, el uso de las grabadoras se había perfeccionado en esa época. En cuanto a esto, Gougenheim (1954: 218) señaló: "L'utilisation de la langue parlée, qui aurait paru chimérique il y a moins de cinquante ans, est devenue possible et même aisée grâce aux progres des magnétophones". De esta manera, el francés fundamental, por un lado, contribuyó y modernizó la metodología de selección y composición de los corpus de estudios, adecuándola a los nuevos tiempos; y, por otro lado, la lista de frecuencia contrastó con las de trabajos previos realizados sobre el francés, como el de Vander Beke (1935).

Respecto a cuántas palabras debía contener la obra, concordaron que –si bien alrededor de 800 palabras era ideal para configurar un francés inicial, que pudiera ir aumentando a medida que se iba pasando de los niveles básico a los avanzados de la lengua– unas 1000 palabras aproximadamente cumplirían con los objetivos del proyecto. En este sentido, bajo el criterio de seleccionar solo aquellos vocablos con una frecuencia igual o mayor a 20, se escogieron 1063 vocablos. Estos fueron expuestos en dos listas, a saber:

- En la primera, se presentaron las palabras organizadas de forma decreciente según la frecuencia y la distribución, siendo la primera palabra el verbo *être* con una frecuencia de 14 083, repartido en los 163 textos del corpus, mientras que la última palabra fue el adjetivo *inventeur*, con una frecuencia de 20, distribuido en 3 textos.
- En la segunda, se expusieron las mismas 1063 palabras, pero en orden alfabético con su número de orden correspondiente de la primera lista, distribución y frecuencia. En ambas listas, se identificó un conjunto de palabras con una equis (x), que indicaba que, si bien habían tenido una frecuencia igual o mayor a 20, serían excluidas de los materiales por razones justificadas.

Según puede observarse en la lista de frecuencias y en la exposición realizada por Gougenheim *et al.* (1964: 114), sacando las palabras gramaticales y los verbos *être* y *avoir*, el primer verbo en aparecer fue *faire*, cuya frecuencia fue = 3174 y la distribución (dist) 162. Luego, aparecieron los verbos *dire* (f = 2391, dist: 160) y *aller* (f = 1876, dist: 161). Por su parte, los primeros sustantivos en las listas fueron *heure* (f = 545, dist: 117), *jour* (f = 538, dist: 132) y *chose* (f = 477, dist: 121). Y los primeros adjetivos fueron *petit* (f = 863, dist: 143) y *grand* (f = 428, dist: 118).

De esta descripción, se desprenden las siguientes conclusiones. En primer lugar, hay una relación inversamente proporcional entre los porcentajes de las palabras gramaticales y las de contenido léxico-semántico, como los sustantivos, verbos y adjetivos. A mayor frecuencia de palabras gramaticales menor la de sustantivos, mientras que a mayor frecuencia de sustantivos menos

frecuencia de palabras gramaticales. En segundo lugar, las palabras gramaticales pasaron a ser el 90,9% en el primer corte de frecuencias (*au-dessus de* 5001) a 5,7% en el último corte (26 à 21). En tercer lugar, los sustantivos pasaron de 0% en el primer corte a 53,9% en el último. Hay que acotar que los primeros sustantivos (2 únicamente) aparecieron en el cuarto corte de frecuencia. Así pues, entre las palabras con frecuencia de 501 a 1000.

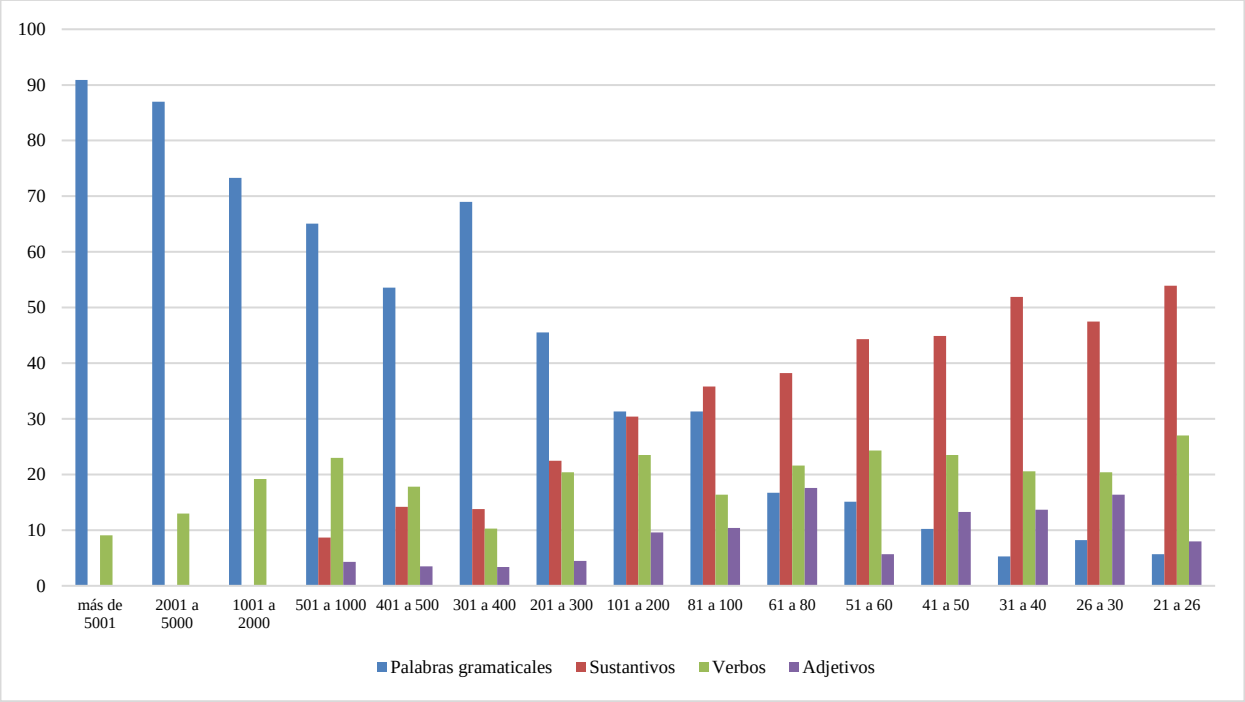
Tabla 1. Las 10 palabras más y menos frecuentes del Francés fundamental.

N° orden	Palabra	Distribución	Frecuencia
1	être (verbe)	163	14.083
2	avoir	163	11.552
3	de	163	10.503
4	je	162	7.905
5	il (s)	160	7.515
6	ce (pronom)	163	6.846
7	la (article)	163	5.374
8	pas (négation)	158	5.308
9	à (préposition)	163	5.236
10	et	161	5.082
1054	éducation	11	20
1055	garage	11	20
1056	période	11	20
1057	pointe	11	20
1058	tableau	10	20
1059	programme	9	20
1060	château	8	20
1061	excursion	5	20
1062	guide	3	20
1063	inventeur	3	20

Fuente: Elaboración propia a partir de Gougenheim *et al.* (1964)

En la Tabla 1 (de elaboración propia, con base en Gougenheim *et al.* (1964)) se muestran las diez palabras más frecuentes del francés fundamental y las últimas diez palabras de la lista; de manera que puedan contrastarse las categorías gramaticales a las que pertenecen las palabras de los primeros lugares con las de los últimos. Y en Gráfico 1, se exponen los porcentajes de los tipos de palabras según los cortes de frecuencias en el Francés elemental.

Gráfico 1. Distribución de los porcentajes de tipos de palabras del Francés fundamental.

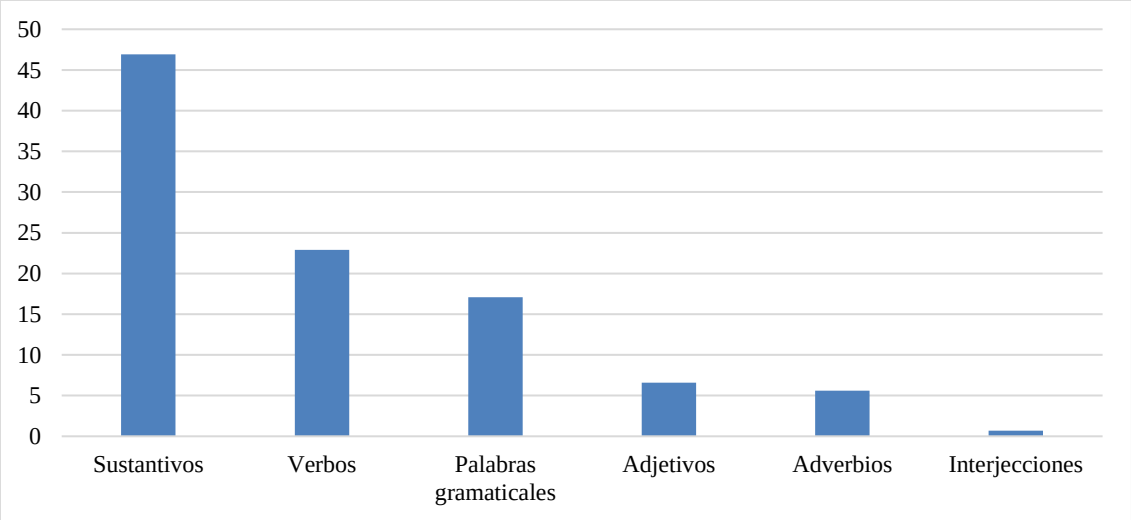


Fuente: Elaboración propia a partir de Gougenheim *et al.* (1964: 115)

En síntesis, puede indicarse que las palabras gramaticales (determinantes, preposiciones, conjunciones, pronombres) y adverbios, ocupan casi exclusivamente los primeros tres cortes de frecuencias, con una ocurrencia que va desde 1001 a 14 083. Por su parte, los sustantivos presentan los mayores porcentajes en los últimos cortes, observándose un predominio a partir del corte 81 a 100. En cuanto a los adjetivos, tienen los mayores porcentajes en los últimos cortes, siendo más alto entre las bandas de 26 a 80 ocurrencias. Por último, los verbos están distribuidos irregularmente en toda la escala, con mayor o menor número de casos, según sus características semánticas.

Directamente sobre el vocabulario del francés fundamental, este quedó constituido por 1475 lemas, siendo 692, el 46,9 %, sustantivos; seguido por 339 verbos, 22,9 %; 253 palabras gramaticales, 17,10 %; los adjetivos fueron 98, 6,6 %; 83 adverbios, 5,6 %; por último, hay que señalar que se incluyeron 10 interjecciones, que se consideraron pertinente, por ser el francés fundamental un vocabulario para la conversación. En el gráfico 2, a continuación, se presenta la distribución de las categorías gramaticales de las palabras del francés fundamental de Gougenheim *et al.* (1964).

Gráfico 2. Distribución porcentual de los tipos de palabras del Francés fundamental.



Fuente: Elaboración propia a partir de Gougenheim *et al.* (1964).

En síntesis, en este apartado se ha querido reseñar la conformación del Francés fundamental; conocer: i) el contexto en el que se iniciaron sus trabajos, ii) los aspectos metodológicos, y iii) los resultados generales. De esta revisión se desprende que, en primer lugar, el Francés fundamental además de ser concebido como una herramienta útil para la enseñanza de la lengua francesa, también cumplió un importante papel político en la expansión y posicionamiento de dicho idioma en el mundo, sobre todo, en un contexto en el que el inglés estaba ganando terreno como lengua franca. Asimismo, los lexicógrafos franceses trataron de solventar algunos de los inconvenientes que se observaron en el inglés básico, especialmente, el referido a la adecuación de un vocabulario lo más cercano posible a la realidad lingüística de los hablantes. En este sentido no se pretendía crear una lengua universal artificial, desvinculada de la conversación natural. Pues, más allá del interés político, apremiaba un verdadero interés lingüístico enfocado en la enseñanza, como apunta Mateo García (1997: 27):

> los intereses de los autores franceses antes mencionados en bastantes aspectos no dejaron de apuntar en esa dirección: conocer mejor los inventarios léxicos propios de las situaciones más habituales en la comunicación ordinaria para llegar, más tarde, a programar de manera más adecuada la enseñanza de lengua.

En segundo lugar, con miras a alcanzar una obra sólida para la enseñanza de lengua, los investigadores franceses replantearon los aspectos metodológicos de la frecuencia léxica. En este sentido, el trabajo de Gougenheim, Michéa, Rivenc y Sauvageot constituyó un gran avance en los estudios léxico estadístico, tanto por la concepción de un nuevo campo investigativo, como la disponibilidad léxica, como en el tratamiento y procesamiento de corpus orales; y en el mejoramiento de los criterios estadísticos para la selección de los vocablos. Sin embargo, los planteamientos teóricos quedaron al debe. Si bien, Michéa tomó en cuenta la capacidad asociativa del léxico basado en la

memoria semántica, no se profundizó, por ejemplo, en: la organización del lexicón y las relaciones semánticas entre las palabras (cf. Ávila Muñoz y Sánchez, 2010: 45). Al respecto, debe mencionarse la investigación de Hernández Muñoz (2006) en la que se ofrece una propuesta teórica cognitiva integral acerca de la disponibilidad léxica, que ha resultado un aporte a la concepción teórica y conceptual de este campo disciplinar.

1.5. Trabajos precursores de disponibilidad léxica

Después de haberse adentrado en los pormenores de la elaboración del Francés elemental, resulta imperativo conocer los trabajos precursores sobre léxico disponible. Así pues, en las siguientes páginas se reseñan algunos de estas investigaciones. La primera obra de referencia de disponibilidad léxica en lengua inglesa es *Lexical Availability. A New Aspect of the Lexical Availability of Secondary School Children*, publicada en 1969 por Dimitrijević, la cual es también la primera investigación sobre léxico disponible después del Francés fundamental. El autor toma en cuenta a grandes rasgos la metodología implementada por los lexicólogos franceses, pero agrega algunas adecuaciones metodológicas que repercuten también en aspectos teóricos de la disponibilidad léxica, tales como: i) la reducción del número de centros de interés, ii) el cambio del tiempo requerido para la elaboración de las listas de palabras, iii) la introducción del término "lexical item", y iv) la distinción entre los términos "palabras disponibles" y "palabras parcialmente disponibles" (Mateo García, 1997; Sánchez-Saus, 2011). El lexicógrafo yugoslavo propuso los siguientes centros de interés:

1) *Animals*,
2) *Countryside*
3) *Town*
4) *Enterteinment*
5) *Jobs and professions*
6) *Sciences*
7) *Means of transport*
8) *Politics* (nuevo CI)
9) *Parts of the house*
10) *Food and drink*
11) *Clothes*

Esta reducción del número de áreas temáticas en comparación con el trabajo fundacional de Gougenheim *et al.* (1964) y la selección de este conjunto de campos semánticos estuvo determinado,

primero, por la familiaridad y la representatividad que significaban para los encuestados y, segundo, por la necesidad de no agotar demasiado a los participantes, ya que el tiempo de respuesta por cada CI era 5 minutos. (Mateo García, 1997; Sánchez-Saus, 2011; Santos Díaz, 2015, entre otros).

Considerando los resultados de la investigación y el grado de disponibilidad de las palabras, el autor introduce la distinción entre "palabras disponibles" y "palabras parcialmente disponibles". Las primeras se refieren a las más familiares para el hablante, por lo que el trabajo cognitivo para encontrarlas en el lexicón es mínimo. Este tipo de palabras, entonces, son las que aparecen más y en los primeros puestos de las listas de palabras y, por ende, tienen los índices de disponibilidad más altos (a tenor de los algoritmos estadísticos actuales). Por el contrario, las segundas –palabras parcialmente disponibles– son las que, si bien son conocidas, la tarea de recuperarlas cognitivamente es más difícil, por lo que su número de aparición en los corpus es bajo (Mateo García, 1997; Sánchez-Saus, 2011; Santos Díaz, 2015). Por último, debe mencionarse la introducción del término *lexical item*, que es traducido al español como "unidad léxica". Dimitrijević (1969) lo utiliza para referirse a los casos plurilexicales (ver 1.3.2.) Así, las respuestas como *intestino grueso* e *intestino delgado*, que suelen aparecer en las listas del CI *Partes del cuerpo*, son etiquetadas como unidades léxicas.

En definitiva, Dimitrijević contribuyó con aspectos metodológicos y tangencialmente teóricos que enriquecieron las subsiguientes investigaciones de este campo de estudio. Por una parte, permitió definir la manera de recolectar las listas de palabras, para lo que se basó en poner límite de tiempo para la elaboración de las listas, método neurálgico que sigue utilizándose hoy en día, con la respectiva adecuación del minutero. Igualmente, abrió la posibilidad de considerar nuevos centros de interés, distintos a los del trabajo fundacional en francés, tomando en consideración los objetivos de cada trabajo y las relaciones experienciales de las comunidades de habla analizadas. Por último, dio paso a la reflexión acerca de la terminología utilizada en la léxico-estadística y sus implicaciones metodológicas en la disponibilidad léxica.

Si bien Dimitrijević (1969) adecuó los nombres de los centros de interés aludiendo a razones de familiaridad cultural, no fue sino Mackey (1971) quien demostraría la incidencia de las diferencias culturales en el léxico de las personas. De hecho, Mateo García (1997: 77) afirma que Mackey y su equipo se apegaron a la metodología de Gougenheim *et al.* (1964) con el firme objetivo de "utilizar la disponibilidad como medida de diferencias culturales entre el vocabulario común de Francia y el de Canadá". El trabajo de Mackey (1971) –*Le vocabulaire concret usuel des français et acadiens: étude témoin* (Tomo I) y *Le vocabulaire disponible des enfants acadiens* (Tomo II)– basado en las muestras de niños canadienses resulta importante en los estudios de disponibilidad léxica por varias

razones. En primer lugar, se fundan las bases para los estudios comparativos intermuestrales de DL, pues se contrastan los lexicones de niños franceses y canadienses. En segundo lugar, se proponen lineamientos de lematización, que es una tarea elemental en los trabajos modernos de DL. Contar con una guía de lematización clara permite, no solo desarrollar análisis estadístico adecuado, sino también confrontar dos o más muestras. En tercer lugar, se describen empíricamente las disimilitudes culturales y sociales entre el LD de los participantes canadienses y los franceses. Al respecto, Sánchez-Saus (2011:80) resalta la influencia del contexto cercano de los encuestados en la producción y asociación semántica, lo que demostró que "el entorno influye en la organización léxica de los hablantes". A grandes rasgos, puede decirse que las conclusiones de Mackey motivaron la ampliación del enfoque de la disponibilidad léxica, pasando de la lingüística aplicada a la enseñanza de lengua a disciplinas vinculadas con la sociología y la psicología, y, en el caso específico de las ciencias del lenguaje, a la psicolingüística, sociolingüística y etnolingüística (López Morales, 1995-1996; Mateo García, 1997; Ávila-Muñoz y Sánchez, 2010; Sánchez-Saus, 2011, entre otros). Así pues, gracias a los trabajos de Dimitrijević (1969) y Mackey (1971), se ha llegado a la posibilidad de adecuar el número de centros de interés, según los objetivos de la investigación y las características sociodemográficas y culturales de los encuestados.

No cabe duda de que la influencia del trabajo de Gougenheim y su equipo influyó no solo a investigadores europeos y americanos, sino también a africanos, especialmente, en las antiguas colonias francesas, como Senegal y Camerún, donde se llevaron a cabo estudios de disponibilidad léxica con fines pedagógicos y sociológicos, considerando datos provenientes de una lengua (algunas de las autóctonas africanas) o de dos lenguas distintas (contraste entre alguna lengua africana y el francés, por ejemplo), entre los que pueden mencionarse los dirigidos por el *Centre de Linguistique Appliquée* de Dakar en 1966, y los de Fall (1976), Njock (1978) y Gontier (1986) (Santos Díaz, 2015: 71). Sin embargo, en estas líneas se reseña la pesquisa de Njock (1978), titulado: *L'univers familiar de l'enfant africain*, que resalta, entre otras razones, porque i) se analizaron muestras de sujetos bilingües (basaa y francés); ii) se contrasta el léxico de dos lenguas de troncos y familias distintos, pero que han estado en contacto por mucho tiempo, y iii) ofrece datos sobre rasgos sociales y cognitivos a partir del léxico.

Njock (1978) parte de la premisa de que las palabras disponibles forman parte de una red asociativa, de manera que cuando una idea se activa, emergen las palabras relacionadas con ella. Para llevar a cabo esta investigación, el autor analiza los datos obtenidos de una muestra estratificada de 220 niños bilingües (basaa-francés), cursantes de los niveles básicos de educación. Los centros de

interés propuestos por Njock fueron: 1) *Las partes del cuerpo*; 2) *Comida y bebida*; 3) *Ropa (de hombre y de mujer)*; 4) *Casa: materiales de construcción y muebles*; 5) *Trabajo en el campo*; 6) *Cocina*; 7) *Aula*; 8) *Medios de transporte*; 9) *Música y danza-géneros*; y 10) *Juegos y entretenimiento*.

Con base en los resultados, el investigador afirma que existen ocho factores que inciden sobre el léxico disponible y las asociaciones de ideas en la mente de los sujetos, a saber: 1) los centros de interés, 2) la época en la que se realiza la prueba, 3) la estación del año, 4) el medio socioeconómico, 5) el sexo, 6) el tipo de establecimiento educativo, 7) el nivel educativo o escolaridad, y 8) la edad (Santos Díaz, 2015: 73). En este sentido, los datos mostraron, por un lado, las relaciones existentes entre el caudal léxico y factores de índole social, y, por otro lado, la influencia de factores cognitivos como la experiencia y la corporeidad con el contexto cercano en el que viven y se desarrollan los encuestados.

El primer estudio de disponibilidad léxica en el que se tomó en cuenta el español fue la tesis de maestría de John Bailey Victery, Titulada *A study of Lexical Availability among Monolingual-Bilingual Speakers of Spanish and English*. El autor propuso las siguientes hipótesis: i) el grado de disponibilidad de las palabras difiere, principalmente, por la lengua analizada, y ii) los factores extralingüísticos sexo y geografía inciden en el grado de disponibilidad (Bailey Victery, 1971: 6-7). Para llevar a cabo la investigación, el autor analizó los listados elaborados por 99 estudiantes de secundaria y técnica, distribuidos equitativamente en tres grupos. La prueba de disponibilidad léxica estuvo compuesta por diez centros de interés, que integraban algunos del estudio pionero de Gougenheim *et al.* (1964) y del trabajo de Dimitrijević (1969). Estos eran los mismos para el español y el inglés, con sus respectivas traducciones a los dos idiomas, como se ilustra en la Tabla 2.

Tabla 2. Centros de interés de la investigación de Bailey Victery (1971)

N°	Disponibilidad léxica en español	Disponibilidad léxica en inglés
1	Animales	Animals
2	La ciudad	The city
3	Diversión	Entertainment
4	Familia	Family
5	Dios	God
6	Trabajos y profesiones	Jobs and professions
7	Medicina	Medicine
8	Música	Music
9	Espacio	Space
10	Guerra	War

Fuente: elaboración propia basada en Bailey Victery (1971)

Bailey Victery se basó en la propuesta metodológica de Dimitrijević (1969), con algunas adecuaciones, de las que debe resaltarse la estipulación de dos minutos límites para la realización de

las listas de palabras. Es decir, la primera investigación en la que se estableció el requisito de dos minutos, por centro de interés, como tiempo límite para que los encuestados escriban libremente todas las palabras. Este aspecto repercutió de tal manera en la comunidad científica que ha sido acordado en las metodologías de las subsiguientes investigaciones, como la de Mena Osorio (1986), quien corroboró la idoneidad de este requisito. En efecto, dada la factibilidad de los dos minutos para desarrollar cada test de DL, el PPHLD lo ha planteado dentro de su diseño, lo mismo en la presente tesis doctoral.

Como se ha constatado en los párrafos previos, el primer estudio de disponibilidad léxica que tomó en cuenta la lengua española fue la tesis de maestría de Bailey Victery (1971). Sin embargo, como apunta los autores del área (cf. Mateo García, 1997:79; Gómez Devís, 2004: 13, Herranz, 2020: 22, entre otros), fue la investigación de López Morales (1973), y sus posteriores aportaciones (1978, 1979, 1986), las que impulsaron el interés de los lingüistas en el mundo hispánico por analizar el léxico disponible de las distintas comunidades de habla españolas de ambas orillas del Atlántico, donde se ha aprovechado al máximo (Sánchez-Saus, 2011).

En 1973, López Morales presenta su estudio titulado *Disponibilidad léxica en escolares de San Juan*, que seguía a grandes rasgos las pautas de los trabajos de Gougenheim *et al.* (1954, 1964) y de Mackey (1971). El objetivo de esta investigación era examinar el caudal léxico de escolares puertorriqueños desde la vertiente sociolingüística. Para esto, el autor analizó el léxico disponible de 63 escolares de San Juan de Puerto Rico, provenientes de tres escuelas de dos niveles socioeconómicos. Las áreas nocionales consideradas fueron 10, a saber: 1) Alimentos, 2) Animales, 3) La casa, 4) La cocina, 5) El cuerpo humano, 6) Materiales de construcción, 7) Muebles, 8) Naturaleza, 9) La ropa, y 10) Transporte. En años posteriores, el autor amplió la muestra, aplicando las encuestas a 558 escolares de todo el país; los resultados fueron publicados en el libro titulado *Léxico disponible de Puerto Rico* (1999).

En un trabajo de 1978, López Morales realiza un estudio en el que se enfoca en analizar el componente léxico de la lengua española, considerando sus características estadísticas. Para esto, toma en cuenta los resultados de su estudio de disponibilidad léxica de 1973 y los de frecuencia léxica de Rodríguez Bou de 1952. Con esta investigación, el autor intenta aportar una descripción del léxico de Puerto Rico al considerar dos tipos de listados de palabras: por un lado, las de léxico disponible y, por otro lado, las de frecuencia léxica, relacionándola con factores sociales, como la edad de los informantes y el nivel educativo. Si bien, la reflexión central giraba en torno a la incidencia de variables sociales sobre el léxico, López Morales también se interesó en la programación del

vocabulario en la escuela, que era un contexto variable (Mateo García, 1997: 80). Entre sus conclusiones, el lingüista señala que existen divergencias en el léxico de ambos trabajos, debido al tiempo transcurrido en la recogida de los datos en uno y otro estudio, a los factores sociales de los informantes, y, sobre todo, al hecho de que cada uno se desarrolló con procedimientos metodológicos distintos (Mateo García, 1997: 80; Santos Díaz, 2020: 48).

Un año más tarde, en 1979, López Morales aborda el problema de la *hipótesis del déficit lingüístico* propuesta por Bernstein, más con la intención de aportar datos empíricos que de adentrarse en una discusión o reflexión teórica sobre la propuesta bernsteiniana, tal como indican Mateo García (1997: 80) y Herranz (2020: 23). Así pues, el autor tomó en cuenta el nivel léxico de la lengua para determinar la incidencia de los factores sociales, como el nivel socioeducativo, en el desarrollo cognitivo de los escolares, tal como apunta la teoría del sociólogo inglés. Basándose en la metodología de la disponibilidad léxica, López Morales analizó el léxico disponible de escolares de San Juan de Puerto Rico, cursante de los niveles 1. °, 3. ° y 5. ° grado de escuela elemental pública; los informantes pertenecían a dos niveles socioeconómicos: bajo y medio; igualmente, el lingüista consideró el sexo de los niños. Los centros de interés aplicados fueron los utilizados en los primeros trabajos de este campo de investigación, tomados de los estudios de Gougenheim y compañía. El análisis de los casos determinó que el nivel socioeconómico incide en el léxico, puesto que pudo apreciarse una clara diferencia en el léxico de los escolares de nivel bajo y los de nivel medio. Respecto a la teoría sociológica de Bernstein, López Morales indica que esta por sí misma no puede dar luces sobre las diferencias sociales, por el contrario, necesita de otras herramientas, como la de la disponibilidad léxica para ser corroborada o rechazada.

Bajo el cuestionamiento de si la posición de las palabras en las listas podría conllevar variaciones en los distintos estudios o no (Trigo Ibáñez, 2011: 25), en 1983, López Morales y Loran Santos revolucionaron los estudios de disponibilidad léxica al crear una fórmula estadística compleja que permitiera determinar correctamente las palabras más disponibles. Dicha fórmula[7] buscaba evaluar la frecuencia, la espontaneidad y la disponibilidad; y estaba basada en la ponderación de la frecuencia y la posición de los vocablos en las listas de palabras. De esta manera, se afianzó la concepción de una nueva línea de investigación o terreno dentro de las ciencias del lenguaje: la lingüística estadística, tal como ya había sido postulada y propuesta por Guiraud (1960) y Müller (1973).

[7] Para más detalles acerca de las fórmulas estadísticas para el cálculo de disponibilidad léxica, diríjase a la sección 1.7.

A su vez, López Morales no solo asumió la disponibilidad léxica desde el enfoque de la sociolingüística al correlacionar el sexo y nivel socioeconómico con el caudal léxico, sino también por ser uno de los primeros (si no el primero) en tomar en cuenta la edad de los informantes: específicamente se centró en aquellos sujetos con edades cercanas a los 18 años, la población joven que estaba cursando los últimos años de secundaria (o los había recién terminado) y no había ingresado aún en la universidad. Esto con un doble propósito, el primero era conocer el léxico adulto de los puertorriqueños, mientras que el segundo era analizar un léxico no contaminado por los tecnicismos de los estudios universitarios. Además, el investigador sugirió que, primero, se utilizasen los dieciséis centros de interés propuestos por Gougenheim *et al.* (1964); segundo, se dejase las listas abiertas de palabras; por último, se establecieran dos minutos como el tiempo requerido para la elaboración de las listas de palabras por cada centro de interés.

1.6. Proyecto Panhispánico de Léxico Disponible

Como se ha señalado, en los primeros años de la DL, los investigadores siguieron a grandes rasgos la metodología implementada por los precursores franceses, la cual fue adecuándose según los intereses científicos particulares. A pesar de que los estudios compartían una misma base epistemológica, existían diferencias que impedían desarrollar trabajos contrastivos, de las que pueden señalarse: i) los centros de interés analizados, ii) la cantidad de ellos aplicados en cada proyecto, iii) el tiempo para que los encuestados escribieran sus respuestas, iv) los sujetos encuestados y, particularmente, v) la edición o lematización de los materiales. En este contexto, debe recalcarse la contribución que Mackey (1971), junto a su equipo, realizaron –gracias a las pautas generales de lematización– para intentar solventar dichas disparidades entre las pesquisas y, poder así, desarrollar comparaciones intermuestrales. El especialista partía de la idea de que al contar con un modelo común de edición que fuese aplicado en los diversos proyectos lexicográficos de DL, se dispondría de datos comparables en este campo. En efecto, el lexicógrafo y su grupo, una vez que lograron homogeneizar su corpus y el de Gougenheim *et al.* (1964), pudieron relacionar el LD de Canadá y Francia. Relativo a lo anterior, Samper Padilla (1998b: 311) señala que:

> De acuerdo con Mackey, uno de los principios que debe guiar el trabajo de edición de los materiales de cada zona es facilitar, en la mayor medida posible, el cotejo dialectal, una labor completamente necesaria para acercarnos a la delimitación del léxico ‘español general’.

En lengua española, esta preocupación la expresó Humberto López Morales, quien, con el apoyo de la Asociación de Lingüística y Filología de América Latina (ALFAL), propuso la creación de un proyecto internacional con el que pudieran coordinarse diversas investigaciones sobre

disponibilidad léxica en el ámbito hispánico. Hogaño, el Proyecto –cuyo objetivo principal es "Elaborar diccionarios de Disponibilidad Léxica para las diversas zonas del mundo hispánico"[8] (Dispolex, 2023)– es un referente para los diferentes grupos interesados en la disponibilidad léxica; paralelamente, es una bisagra que los vincula a pesar de las distancias.

Al plantear una metodología común, el PPHLD ofrece la ventaja de unificar las diversas tareas, tales como: i) la recolección de los datos, ii) la edición de los materiales, iii) la selección de los participantes, entre otras (Moreno Fernández, 2012; Paredes, 2014). De hecho, en la página web oficial del Proyecto se detalla que: "La homogeneidad de criterios permitirá establecer comparaciones de tipo lingüístico, etnográfico y cultural, dibujar áreas de difusión y, en general, servirá de punto de partida para análisis posteriores" (Dispolex, 2023). A grandes rasgos, siguiendo a Samper Padilla (1998b), Bartol (2006), Saralegui y Tabernero (2008) y Sanchéz-Saus (2011), los puntos planteados en la metodología del PPHLD son:

- Las muestras deben estar basadas en las respuestas de hablantes nativos del idioma, quienes se encuentren cursando estudios de secundaria o preuniversitarios. La selección de esta población se debe a que se considera a estos informantes como lingüísticamente adultos, pero "sin una especialización universitaria, técnica o profesional que pueda "contaminar" la representatividad de un hablante medio que sigue la norma general de la comunidad" (Saralegui y Tabernero, 2008: 746).
- La utilización de dieciséis centros de interés, que concuerdan *grosso modo* con los usados por Gougenheim *et al.* (1964).
- Los informantes escriben sus listas de palabras en cuadernillos de papel.
- El tiempo de respuestas por cada área nocional es dos minutos.
- Utilización de las variables extralingüísticas: sexo, nivel sociocultural, tipo de centro educativo y zona geográfica.
- El empleo de los mismos criterios de edición de los datos y posterior procesamiento.

En las dos orillas del Atlántico se cuentan equipos de investigación enfocados en el LD, los cuales, según Bartol (2006), pueden organizarse en tres grupos: i) los que han adoptado las directrices del PPHLD, ii) los que han asumido parcialmente las pautas del Proyecto, y iii) los que analizan el LD aprendientes de español como segunda lengua. De manera particular, a continuación, se mencionan algunos de los trabajos de los puntos i) y ii).

[8] Para más detalles acerca del PPHLD, ver: http://www.dispolex.com/info/el-proyecto-panhispanico

Algunas de las contribuciones relativas al léxico potencial de España, en el marco del PPHLD, son: Samper Padilla (1998a), Samper Padilla y Hernández (1997), Palmas de Gran Canaria; Ayora (2004), Bartol (2004) y Mateo (1998), sobre Ceuta, Soria y Almería, respectivamente. También se hallan las de Gómez Devís (2004), en Valencia; Ahumada (2006), en Jaén; Ávila Muñoz (2006), en Málaga; Fernández Juncal (2008), en Burgos; Trigo Ibáñez (2011), en Sevilla; Casanova (2017), en Castellón, entre otros. Por su parte, en el territorio americano, la distribución del número de investigaciones es discontinua; pues, mientras en algunos países se cuentan con publicaciones al respecto, a saber: Alba (1995), República Dominicana; López Morales (1999), Puerto Rico; Bonilla (2004), Mateus-Ferro y Williams (2006), Garzón y Penagos (2016), Henríquez, Mahecha y Mateus-Ferro (2016), Colombia; Echeverría *et al.* (1987), Valencia y Echeverría (1999), Valencia (2010) Chile; Pacheco *et al.* (2017), Cuba; Saine Camargo (2008) y Wingeyer (2014), Argentina.

En cuanto a los estudios que han dado cuenta sobre el léxico disponible a través de métodos parcialmente similares a los del PPHLD, pueden enumerarse las de Justo Hernández (1986), Cañizal Arévalo (1987), Ruiz Basto (1987), López Chávez (1994), en México; Murillo (1993, 1994, 1998), Sánchez y Murillo (1993), Costa Rica; y Azurmendi (1986), España. Estos se han distinguido porque han utilizado centros de interés diferentes o el número de ellos es disímil o las edades de los sujetos no concuerdan con la de estudiantes de secundaria.

En este punto, de manera especial, deben reseñarse algunos de los aportes que los lingüistas y especialistas chilenos han ofrecido desde este bloque. Pues, a raíz de la revisión de los títulos de las diversas publicaciones, se observa que el foco de la DL se dirige hacia estudiantes universitarios (cf. Fasces *et al.*, 2009; Cerda *et al.*, 2017; Blanco *et al.*, 2020), vocabulario sobre matemática (cf. Salcedo *et al.*, 2013; Ferreira *et al.*, 2014; Salcedo *et al.*, 2017, entre otros), comprensión lectora en niños (Cepeda *et al.*., 2017), vocabulario de adultos mayores (Urzúa, 2018), literacidad en salud (Castillo Fadić y Santos Díaz, 2021) y en la elaboración de aplicaciones para los análisis léxico-estadísticos, como Dispogen (Echeverría *et al.*, 2005) y Dispografo (Echeverría *et al.*, 2008).

En virtud de lo anterior, la presente tesis doctoral se incluye en este segundo conjunto de estudios, puesto que, si bien la base metodológica tiene relación con la del Proyecto, se distancia de este, en virtud de que, por un lado, los sujetos son universitarios y, por ende, han entrado en contacto con tecnicismos; por otro lado, únicamente se han usado cuatro de los dieciséis ejes temáticos; y, por último, una de las tres muestras analizadas se recogió a través de una página web *ad hoc*.

En definitiva, esta somera reseña del PPHLD demuestra el interés que acarrea este tipo de estudios en la comunidad lingüística hispánica, desde distintos enfoques y disciplinas. En el próximo

epígrafe, se da cuenta de la fórmula estadística compleja con la que se calcula el índice de disponibilidad.

1.7. Fórmula para el cálculo del índice de la disponibilidad léxica

La revisión bibliográfica da cuenta de la evolución que han tenido los planteamientos epistemológicos de la disponibilidad léxica a partir de los trabajos léxico-métricos basados en la frecuencia. Así pues, no es ilógico que, en paralelo, se hayan adecuado también los procedimientos analíticos y las fórmulas matemáticas para el cálculo del léxico disponible. En efecto, a partir de las debidas revisiones metodológicas, el equipo de lexicógrafos franceses organizó los vocablos más disponibles en función del número de menciones en el corpus, sin tomar en cuenta la cantidad de veces que fuese escrito por un mismo informante. Dicho de otra manera, la ocurrencia o frecuencia del lema disponible X coincidía con la cantidad de sujetos que lo reportaban, por lo tanto, a mayor número de menciones, mayor era la posición del lexema en el listado del vocabulario disponible (Ávila Muñoz, 2010). Según Callealta y Gallego (2016), esta metodología fue replicada en algunos de los estudios pioneros del campo, a saber: Mackey (1971), Dimitrijević (1969), López Morales (1973) y Azurmendi Ayerbe (1983).

Sin embargo, como reseña Santos Díaz (2015: 103), Njock (1978) estableció "la distinción entre el índice de frecuencia y el rango asociativo de la palabra", a partir de esta el autor inició sus reflexiones acerca de la relevancia que deben poseer las relaciones asociativa de las piezas léxicas en la determinación de los vocabularios disponibles. Con base en sus reflexiones, postuló una fórmula para el cálculo de rango asociativo, en la que se tomaba en consideración: i) el rango medio de la palabra dentro de un centro de interés, ii) la suma de los rangos individuales de la lexía y iii) el número de participantes que escribió el lema. No obstante, sobre la base de los resultados comparativos entre el francés y el basaa, el lexicógrafo señaló que la disponibilidad y su índice miden aspectos distintos del léxico (Santos Díaz, 2015).

En la otra orilla del Atlántico, López Morales –considerando lo poco afinada que resultaba la metodología inicial de Gougenheim *et al.* (1964) – trabajó junto a Lorán en la creación de una fórmula matemática que fuera capaz de deducir la disponibilidad de una palabra, en la que se combinaran la frecuencia y la posición en las listas. Así pues, ambos investigadores desarrollaron dos ecuaciones en las que integraron un factor de ponderación decreciente de tipo potencial, cuyo valor era menor que 1, representado como λ^{i-1}; debe señalarse que en los primeros trabajos $\lambda = 0{,}90$. Con esta, lograron definir un valor para la jerarquización de los vocablos o *índice de disponibilidad léxica* (en su sigla, IDL). El IDL es el más importante en los estudios de este tipo, puesto que indica el grado de

disponibilidad que tiene una palabra para un grupo de sujetos en un tópico específico (cf. Valencia, 1997; Ávila Muñoz, 2010, 2016; Ávila Muñoz y Sánchez, 2011). Empero, los investigadores puertorriqueños detectaron que el tamaño muestral incidía en la suputación, por lo que concibieron dos modelos matemáticos que solventaran dicha cuestión (Lorán y López Morales, 1983; Callealta y Gallego, 2016), estos se ilustran en (2a) y (2b), seguidamente:

(2a)

$$\text{d(p)} = \sum_{i=1}^{n} i - 1\, x_{pi} \quad \text{siendo} \quad x_{pi} = \frac{f_{pi}}{N_i} \qquad \textit{versus}$$

(2b)

$$\text{d(p)} = \sum_{i=1}^{n} i - 1\, \frac{f_{pi}}{N_i}$$

En donde:
n = máxima posición alcanzada por las palabras;
x_{pi} = frecuencia relativa de la palabra p en la posición i;
f_{pi} = frecuencia absoluta de la palabra p en la posición i;
N_i = número de informantes cuyas listas contenían la posición i.

Estas contribuciones de Lorán y López Morales (1983) se utilizaron en los estudios de Román (1985), Mena Osorio (1986), Justo Hernández (1986), Echeverría *et al.* (1987) y Cañizal Arévalo (1987). Sin embargo, se observó que la respectiva fórmula aplicada presentaba el problema de que –a partir de cierta posición de una palabra X en la lista, particularmente, el puesto veintitrés– asignaba un coeficiente de mayor valor al que realmente se ajustaba a la posición de la lexía, lo que suscitaba que la pieza léxica X obtuviera un índice irregular (Callealta y Gallego, 2016: 42). En este contexto, Butrón (1991) se propuso postular una manera de ordenar las unidades léxicas de las "listas-respuestas", a través de un conjunto de artificios matemáticos que pudieran enmendar las fallas habidas en las ecuaciones previas. Su punto de partida fue crear un ordenamiento "entre centros de interés a base de las palabras provistas por un grupo de informantes", al que denominó *índice del centro de interés para el grupo* (Butrón, 1991: 81). Para lo cual, tomó en cuentan las siguientes propiedades:

- La variedad de palabras diferentes
- La frecuencia u ocurrencia
- La posición dentro de las listas

Al final, la lingüista puertorriqueña logró aportar más de cuatro operadores matemáticos, sólidamente razonados, para el ordenamiento de las unidades léxicas disponibles por centro de interés. Pese a esto, no pudo repararse totalmente el desajuste de los modelos inferenciales de Lorán y López Morales (1983), empero sí alcanzaba a controlarse.

En este contexto, López Chávez y Strassburger Frías, en 1987, consiguieron adecuar la fórmula Lorán-López Morales (1983), al plantear que la posición en la que aparece un vocablo debía ser elevada a un exponente complejo que ofreciera valores asintóticos a cero. Los especialistas mexicanos presentaron una nueva fórmula estadística –a la que integraron el número de Euler (*E*, también conocido como constante de Napier) elevado al exponente -2.3, verdadero responsable de la ponderación (López Chávez, 1992: 33) –, con la que se cuantificaba eficientemente el índice de disponibilidad léxica. Con esta medida, López Chávez y Strassburger Frías (1987) sortearon los escollos que manifestaban los artificios matemáticos hasta ese momento propuestos (cf. Urzúa, Sáez y Echeverría, 2006; Ávila Muñoz y Sánchez, 2011; Callealta y Gallego, 2016: 44). Sin embargo, López Chávez y Strassburger Frías presentaron en 1991, una adecuación de la fórmula de 1987. En la Figura 6, se especifican las características de esta ecuación:

Figura 6. Fórmula López Chávez y Strassburger Frías (1987, 1991)

$$D(P_j) = \sum_{i=1}^{n} e^{-2.3} \left(\frac{i-1}{n-1}\right) \frac{f_{ji}}{I_i}$$

Donde:
n = máxima posición alcanzada en el CI;
i = número de la posición en cuestión;
j = índice de la palabra tratada;
e = número de Euler o constante de Napier (2,718281828459045…);
f_{pi} = frecuencia absoluta de la palabra j en la posición *i*;
I_1 = número de informantes;
D(Pj) = Disponibilidad de la palabra *j*.

La fórmula de los profesores mexicanos sigue aplicándose actualmente en muchas de las investigaciones de esta área del conocimiento, sobre todo en aquellas vinculadas con el PPHDL. En efecto, algunos de los programas informáticos utilizados la tienen integrada en sus sistemas para el cálculo de la DL. Concretamente, en la Universidad de Concepción, Chile, Echeverría, Urzúa y Figueroa crearon en 2005 Dispogen, que es una aplicación por medio de la cual se calcula no solo el IDL, sino también las frecuencias absolutas y relativas, frecuencia acumulada y porcentajes de las lexías en un centro de interés. Adicionalmente, Dispogen arroja los cálculos de número de palabras y vocablos, y promedios.

En este capítulo se han ilustrado y explicado los enfoques metodológicos desde los que pueden llevarse a cabo los estudios de vocabularios reducidos, enfatizando en el cuantitativo, ya que es desde el cual se abordan las investigaciones de DL. Igualmente, se han expuesto las características de los

diccionarios de frecuencia y léxico básico, con la finalidad de mostrar las disimilitudes con los vocabularios de LD. Además, se reseñaron los hitos de la elaboración del Francés fundamental y los posteriores trabajos realizados a partir de los métodos planteados por los lexicógrafos franceses. Asimismo, se revisó el camino que desembocó en la creación de la fórmula estadística compleja para el cálculo del índice de disponibilidad léxica; así como la conformación del PPHLD. Todo esto a la luz de los conceptos lexicológicos, lexicográficos y, sobre todo, léxico-estadísticos. En virtud de esto, en el capítulo siguiente, se expone el modelo metodológico diseñado para esta tesis doctoral, basado en los postulados teóricos aquí desarrollados.

Capítulo 2. Metodología

> "El método científico es un procedimiento general
> que se sigue para alcanzar el conocimiento científico,
> mientras que las técnicas son procedimientos concretos,
> operativos, que se utilizan en el trabajo científico
> para llevar a cabo las distintas etapas del método"
> (López Morales, 1994b: 18)

Como se ha especificado en los epígrafes previos, esta tesis doctoral se fundamenta en los principios de los estudios de disponibilidad léxica, por lo que, a partir del marco teórico conceptual, así como de la revisión de los antecedentes, se ha diseñado un modelo metodológico que ha permitido abordar los objetivos trazados. En este sentido, Briones (2002: 25) señala que el diseño metodológico se refiere a: "la estrategia que se utiliza para cumplir con los objetivos de esa investigación. En términos prácticos, tal estrategia está compuesta por una serie de decisiones, procedimientos y técnicas que cumplen funciones particulares". Conforme a lo anterior, las finalidades de este capítulo son describir y explicar los pasos y procedimientos –al igual que sus bases epistémicas– llevados a cabo para el análisis del léxico disponible de los grupos de estudiantes de Educación Básica y Letras Hispánicas de la UC. La estructura de esta sección es la siguiente:

En primer lugar, se detalla el tipo de investigación. Segundo, se explica la selección de los informantes y el tamaño de la muestra; esta última se determinó a través del cálculo de potencia estadística mediante el programa G*Power 3. Tercero, se describen las características de los informantes y la conformación del corpus. Sobre este, debe resaltarse que está compuesto por muestras de fuentes escritas en papel y en soporte digital. Cuarto, se presentan los instrumentos y los dos métodos utilizados para la recolección de los datos. Quinto, se explica la manera cómo se realizó la edición de los materiales. Sexto, se reseña el plan utilizado para la codificación de los casos, paso previo a la ejecución de los análisis computacionales.

2.1. Tipo de investigación

De manera general –siguiendo los planteamientos de López Morales (1994: 23-25)–, puede definirse esta investigación en los siguientes términos. Por su objeto de estudio, corresponde al nivel léxico de la lengua, puesto que da cuenta del vocabulario de tres grupos de estudiantes universitarios, quienes cursaban (en el momento de responder los cuestionarios) carreras de los campos del saber de Educación y Humanidades y Artes (UNESCO-UIS, 2013: 76). En cuanto al alcance temporal, es sincrónica, ya que se ha analizado el corpus, según un único período, que abarca desde el segundo semestre del 2020 y el primer semestre del 2022.

Además, los fines analíticos de los textos no se guiaron por la distinción de puntos temporales, como ocurre en las pesquisas de orden diacrónico. No obstante, se describieron las convergencias y divergencias léxicas de los participantes en relación con el nivel de curso en el que se encontraban, cuando realizaron los test; específicamente, si estaban en 1.er o 4. º año del respectivo programa académico. Sin embargo, sí podría señalarse que una parte del análisis fue *grosso modo* en tiempo aparente, en términos laboviano (Labov, 1966, 1972), ya que se compararon los lexicones de los alumnos de recién ingreso con el de los de los discentes próximos a egresar. Con relación a la fuente, este trabajo es de tipo primario, dado que los datos provienen de primera mano: las listas de palabras fueron elaboradas –tanto en escritura en papel como digital– por los propios encuestados exclusivamente para esta monografía.

Sin embargo, con el propósito de describir detalladamente el marco metodológico de este estudio, se han contrastado las nomenclaturas de Kumar (2014) –quien organiza los tipos de investigación científica, en cuanto a los criterios de: i) aplicación, ii) objetivos y iii) modo– y López Morales (1994b), como se detallan a continuación.

a. Criterio de aplicación del estudio

Este corresponde al que López Morales (1994b: 24) denomina *finalidad*. El presente estudio es de tipo puro, puesto que busca ampliar los conocimientos sobre el caudal léxico de dos grupos, a saber: el de estudiantes de la carrera de Educación Básica y el de alumnos de Letras Hispánicas, ambos programas de la Pontificia Universidad Católica de Chile. Por un lado, esta tesis ha evaluado el método que integra un instrumento alternativo para la recolección del corpus, con lo que se ahonda sobre aspectos metodológicos de esta disciplina. Por otro lado, se ha declarado un conjunto de preguntas de investigación, con la intención de guiar la observación del léxico disponible de los grupos en relación con factores sociológicos y socioeducativos, tales como: *sexo*, *carrera, año o nivel de curso, cantidad de libros leídos* y *frecuencia de lectura optativa*. Así pues, se buscaba cooperar con el conocimiento de las características estadísticas del LD de los participantes.

No obstante, no se descartó completamente el carácter aplicado de esta investigación, ya que también se persigue contribuir con la elaboración de un léxico reducido acerca de áreas nocionales ligadas a las mallas curriculares, que fungieran como mecanismos de valoración del componente léxico de los cursantes de programas de las facultades de Educación y Letras.

b. Criterio del objetivo de la investigación

Esta tesis es descriptiva, correlacional y explicativa. Primero, se describe el caudal léxico de los participantes acerca de los centros de interés: i) *la lectura*, ii) *el profesor*, iii) *la educación*, iv) *juegos y distracciones*, v) *la escuela: muebles y materiales*, vi) *habilidades y cualidades docentes*, vii) *partes del cuerpo*, y viii) *comidas y bebidas*. Segundo, se ha correlacionado el vocabulario de los sujetos con descriptores de tipo sociológico y socioeducativo, para luego explicar *grosso modo* los resultados generales observados. Por último, debe resaltarse que en esta tesis se ha indagado el contraste entre el léxico disponible recolectado mediante pruebas en soporte de papel y el recogido en formato digital, a través de una página web *ad hoc*. Además, se compararon los lexicones de grupos de estudiantes de dos carreras distintas: Educación Básica y Letras Hispánicas. Asimismo, se analizaron campos nocionales nuevos: *la lectura*, *el profesor* y *habilidades y cualidades docentes*.

c. Criterio del modo de investigación

Los estudios de disponibilidad léxica –como disciplina inscrita en la léxico-estadística– exhiben un alto componente matemático en el análisis de los datos, por lo que esta tesis es *per se* de carácter cuantitativo. Concretamente, se aplican fórmulas matemáticas complejas para los cálculos de los índices de: i) disponibilidad léxica (IDL), ii) cohesión léxica (IC) y iii) promedio de palabras (PP o $\overline{X}$). Igualmente, se contabilizan los números totales de palabras (NP) y vocablos o unidades léxicas diferentes (NV). Para esto, la DL recurre a herramientas computacionales. En esta pesquisa particularmente se ha recurrido a: Dispogen (Echeverría *et al.*, 2005), que es un programa creado para los estudios de este tipo; y IBM SPSS®, con el que se realizaron los análisis estadísticos, descriptivos e inferenciales, correspondientes.

Sin embargo, más allá de los cómputos, también se han desarrollado análisis cualitativos del caudal léxico de los encuestados. Este paso se llevó a cabo mediante las comparaciones de las veinte palabras más disponibles de cada centro de interés, en relación con los descriptores. Igualmente, se utilizó el programa Dispografo (Echeverría *et al.*, 2008), que está destinado a trazar las representaciones gráficas de las relaciones entre las palabras disponibles por medio nodos y aristas calculadas a través de un algoritmo especializado.

En virtud de lo anterior, tomando en consideración los planteamientos de Kumar (2014: 14), quien señala que: “In extremely simple terms, the mixed methods approach to social research combines two or more methods to collect and analyse data pertaining to the research problem”, esta

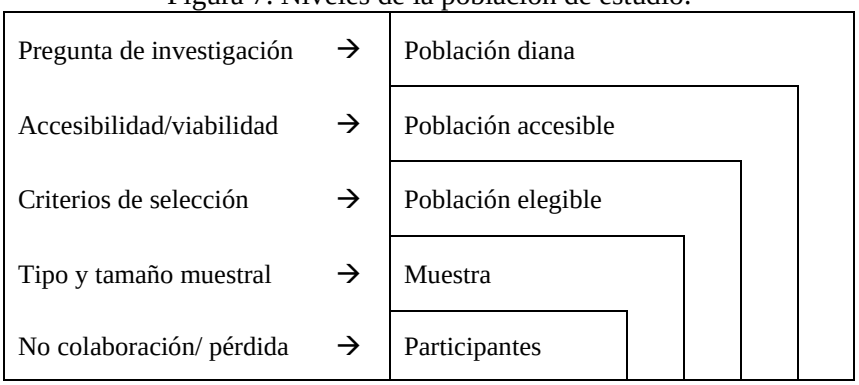
tesis se adecua al carácter mixto del tipo de investigación, puesto que los datos se observaron a través de análisis cuantitativos y cualitativos.

Por último, el tipo de muestreo al que se recurrió para la toma de datos concierne al estratificado al azar, no proporcional. Este, según explica Moreno Fernández (2017: 312), corresponde al que "se aplica el azar, no sobre la población total, sino sobre parcelas de ella, llamadas *cuotas*, que se han determinado de antemano, según los objetivos de la investigación (sexo o género, grupos de edad diferentes, nivel de estudios)". En concreto, antes de llevar a cabo los trabajos de campo, se definieron las características de los informantes y el número de ellos que se necesitaban (según los cálculos *a priori* de G*Power 3 (Faul, Erdfelder, Buchner y Lang, 2007), con el fin de enfocar la recolección del léxico en una población meta específica y abarcable, en este caso: tres grupos del alumnado de Educación Básica y Letras Hispánicas, cursantes de 1.er y 4.º Año, de la PUC. De este conjunto, se colectaron las listas de palabras de los individuos, quienes voluntariamente y al azar, respondieron las encuestas. En el próximo subapartado, se reseña, justamente, este punto referido a la selección de los participantes.

2.2. Selección de los informantes

El corpus de esta tesis quedó compuesto por las listas de palabras escritas por los sujetos que cumplieron con los criterios de selección de la muestra, que se fundamentan en los parámetros de Rebagliato, Ruiz y Arranz (1996). Estos plantearon una organización de los niveles de la población objeto de estudio, los cuales "[…] vienen determinados por factores <sic> interrelacionados pero esencialmente distintos" (Rebagliato *et al.*, 1996: 73). Así pues, el bosquejo se presenta desde la población general hasta los participantes específicos, de quienes, a tenor de la técnica muestral utilizada, se obtendrán los datos (cf. Santos Díaz, 2020: 93). En la Figura 7, se ilustra el esquema de Regabliato *et al.* (1996).

Figura 7. Niveles de la población de estudio.



Fuente: Regabliato *et al.* (1996:74)

- *Población diana*, esta es a la que "pretendemos generalizar los resultados de nuestro estudio" (Regabliato *et al.*, 1996: 73). Esta corresponde al grupo o comunidad total en la que se piensa cuando se planifica un estudio. Siguiendo a Santos Díaz (2020: 93), para llegar a este nivel se ha planteado la siguiente pregunta: ¿qué queremos conocer? De la que se desprende una respuesta general que traza el tema y la población diana: el léxico disponible de estudiantes universitarios de carreras de las áreas de Educación y Humanidades y Artes de una universidad de Santiago que esté acreditada por la CNA.
- *Población accesible o fuente de sujetos*, se refiere a la que está determinada por razones prácticas en función de la accesibilidad o tratamiento directo del investigador con los individuos bajo estudio (Regabliato *et al.*, 1996: 74). Para determinar este grupo poblacional, se han contemplado las siguientes preguntas: ¿cuáles son los posibles informantes abordables por el investigador? ¿Cuál será la vía más idónea para llegar a ellos? Las respuestas a estos cuestionamientos han centrado la investigación en estudiantes universitarios de Educación Básica y Letras Hispánicas de la Pontificia Universidad Católica de Chile. A la par, se ha pensado en las maneras cómo abordar a los sujetos, a saber: primer paso, comunicarse con las directoras y coordinadora académicas de las carreras en cuestión, exponerles los objetivos del trabajo y, justamente, solicitar los permisos necesarios.
- *Población elegible*, concierne a los sujetos que emergen de la población accesible en virtud del cumplimiento de los requisitos específicos de la investigación (Regabliato *et al.*, 1996: 74), para lo cual se ha articulado la siguiente pregunta: ¿a quiénes se les pretende aplicar los cuestionarios de disponibilidad léxica? La respuesta encausó la atención hacia un grupo finito compuesto por alumnos regulares de las carreras antes mencionadas, quienes estuviesen cursando 1.er o 4.º año; en otras palabras, alumnos que hayan ingresado recientemente o que estén próximos a egresar.
- *Muestra*, consiste en un subconjunto estimado de "[…] la población elegible, la frecuencia de una determinada condición o la magnitud de una asociación en dicha población", la cual permite realizar estimaciones sobre todo el conjunto (Regabliato *et al.*, 1996: 74). En relación con los objetivos de esta investigación, se configuraron tres muestras. Dos de ellas están compuestas por las respuestas de los sujetos de la población elegible que contestaron las pruebas por medio de cuestionarios en papel (método o técnica tradicional); y la tercera está constituida por los listados de palabras de los individuos que respondieron las encuestas de manera remota o diferida (método o técnica digital).

Debe acotarse que el número de participantes necesarios para la constitución de las tres muestras se estimó gracias al análisis *a priori* realizado con G*Power 3 (Faul *et al.*, 2007). Este es un programa que favorece la definición de los tamaños muestrales por medio del cálculo de la potencia estadística (Faul, Erdfelder, Buchner y Lang, 2009: 1149). En el subapartado 2.3 se detallan los aspectos de esta y su utilidad en la creación de corpus en las ciencias socias sociales.

- *Participantes*, estos son justamente los sujetos que realizaron efectivamente las pruebas (Regabliato *et al.*, 1996: 74). Este número dista de la población elegible, debido a razones externas a la investigación, tales como: disposición personal, lugar, economía, practicidad, etc.

Particularmente, el estallido social de 2019, por un lado, y la pandemia del covid-19 que llegó a Chile en marzo de 2020, por otro lado, imposibilitaron los trabajos de campo para desarrollar esta investigación, ya que no se podía recoger datos de la manera tradicional, debido a la inaccesibilidad a la población elegible. En virtud de esto, se invirtieron horas en el diseño de un nuevo método con el que se pudieran recolectar las listas de palabras, que siguiera a grandes rasgos las pautas metodológicas del PPHLD. Además, la nueva técnica debía ser, sobre todo, confiable. Ergo, se extendió el tiempo de recogida y construcción del corpus: desde el segundo semestre del 2020 y el primer semestre del 2022. A pesar de las complicaciones externas, se logró crear una técnica que llevó a la construcción de un prototipo digital que era capaz de tomar datos léxicos sin diferir completamente de la metodología del PPHLD.

Una vez superados los problemas externos, ya sea por tener una página web para colectar datos léxicos o ya sea por la flexibilización de las cuarentenas y regreso a las aulas, se consiguió llegar a una parte de la población elegible y, por consiguiente, se logró levantar un corpus basado en las respuestas de 264 encuestados, de los cuales 176 realizaron las pruebas en formato papel y 88, en soporte digital.

A manera de recapitulación de las ideas puntualizadas en la Introducción, la elección de los participantes se debió a que, en primer lugar, las plaza de profesorado representan un campo de alta demanda laboral, por lo que puede despertar el interés de especialistas de carreras distintas al área de la didáctica, por ejemplo: Letras Hispánicas. Los egresados de Letras pueden cursar el componente docente con el fin de incorporarse a dicho sector. Segundo, el ingreso a un puesto académico es cada vez más exigente, por lo que los formando en pedagogía deben contar con habilidades fuertemente consolidadas, entre las que destacan las lingüísticas-discursivas. Por último, la lectura sigue siendo un componente esencial en el proceso de enseñanza-aprendizaje en todos los niveles del sistema

educativo, por lo que es esencial contar con las herramientas necesarias para su promoción y abordaje, desde distintas perspectivas, tanto pedagógicas como lingüística-literarias. Entonces, conviene conocer y contrastar las expresiones léxicas que se tienen los universitarios en formación sobre ejes temáticos diversos. En atención a lo anterior, los universitarios de Educación Básica y Letras Hispánica de la PUC constituyen una población apta para acometer estos problemas.

Una vez que se han descrito las muestras del estudio, respecto a la población; conviene definir el tamaño muestral, el que se explicará a continuación.

2.3. Tamaño de la muestra

Uno de los principales planteamientos en los diversos estudios lingüísticos concierne al tamaño de la muestra. Este tema no está exento de problemas epistémicos, puesto que no hay una respuesta unánime acerca de cuál es el número adecuado de informantes o datos en el que deban fundamentarse los análisis. No obstante, la mayoría de los especialistas converge en que no puede realizarse una investigación basada en toda la población, ya que no resulta factible (cf. Martínez-Lara, 2016; Santos Díaz, 2020). Al respecto, López Morales (1994b: 41) afirma que: "[…] es imposible trabajar con técnicas de censo, es decir, obtener información de todos y cada uno de los individuos que integran la población […] el lingüista tiene que efectuar su análisis sobre material producido por una muestra extraída del universo". Por este motivo, debe seleccionarse una muestra que sea representativa del universo absoluto; confiable y que, por ende, responda las preguntas y objetivos de investigación (Bolívar, 2013).

A grandes rasgos, una muestra es: "un subconjunto de los elementos de una población" (Herrera, Martínez y Amengual, 2011: 28). En tanto que López Morales (1994b: 41) la define como una parte del universo relativo, literalmente: "[…] es el que se obtiene una vez que hemos eliminado a aquellos sujetos que no forman parte de nuestro estudio". Es decir, a partir de un universo absoluto, compuesto por todos los individuos a razón de un censo, se seleccionan solo aquellos que cumplan los criterios pertinentes para la pesquisa. Concretamente, el universo relativo emerge de la conjunción de factores como: sexo, edad, nivel de escolaridad, nivel socioeconómico, entre otros.

En el ámbito de la lingüística –específicamente de la sociolingüística– se han propuesto distintos patrones para intentar fijar el tamaño muestral o "universo relativo". Por ejemplo, Labov (1966) señala que se requiere de un 0,0025 % de la población, cuando está consta de unos 100 000 sujetos; mientras que Sankoff (1978) propone un número exacto, a saber: 150 individuos. En el Proyecto Para el Estudio Sociolingüístico del Español de España y América (PRESEEA), si bien se

toma en cuenta un número fijo de 108 sujetos, la muestra se precisa en función del principio de muestreo aleatorio por cuotas (Moreno Fernández, 2017: 311; Preseea, 2023).

Según Santos Díaz (2020: 94), el tamaño muestral dependerá de diversos aspectos prácticos, entre los que resalta la población, las características del muestreo y la representatividad. En relación con esto, el establecimiento de la muestra puede considerarse a partir de los tipos de análisis cuantitativos que se vayan a desarrollar, puesto que la literatura sustenta que cada estadístico debe contar con un número mínimo de datos. En consonancia con lo anterior, Herrera *et al.* (2011: 28) sostienen que para el cálculo de inferencias es necesario contar con al menos 30 sujetos por grupo. Sin embargo, otros autores señalan que este cómputo es válido solo para análisis simples, mientras que la cantidad debe subir a 50 individuos para cálculos inferenciales significativos (Santos Díaz, 2015).

En resumen, algunas disciplinas aún mantienen abierta la discusión sobre los fundamentos que deben regir el diseño y posterior construcción de los corpus de estudios. Sin embargo, todas concuerdan en que, sea cual sea el principio regulador, este debe estar sólidamente argumentado y en concordancia con los objetivos y la naturaleza del problema de investigación (López Morales, 1994b; Briones, 2002; Mafokosi, 2009; Bolívar, 2013, Moreno Fernández, 2017, entre otros). No obstante, a pesar de este aparente relativismo existente en algunas corrientes para la determinación del N muestral; en otras especialidades –sobre todo dentro de las ciencias sociales y biomédicas, por ejemplo: psicología– se aboga por el empleo del cálculo *a priori* de la potencia estadística, como forma confiable de esbozar el corpus. En esta misma línea, se propugna también realizar los análisis *post hoc* de potencia a fin de advertir la fiabilidad de los estadísticos desarrollados con el N muestral utilizado.

2.3.1. Definición del tamaño de la muestra mediante el cálculo de la potencia estadística

A fuer de lo anterior, desde la psicología, se ha venido recomendando que, en los diseños metodológicos, se informe la potencia estadística, puesto que el valor del *poder* –término que también se le acuña– permite comprobar la validez y eficacia de los resultados cuantitativos inferenciales. Además, este tipo de análisis afianza la credibilidad de los hallazgos. De forma precisa, la potencia estadística es un índice que se refiere “al grado de probabilidad de rechazar una hipótesis nula cuando esta es realmente falsa, es decir, a la capacidad de una prueba para detectar diferencias entre grupos cuando estas están presentes” (Cárdenas y Arancibia, 2014: 211-212).

En líneas generales, la potencia resulta de la resta del promedio de error Tipo II (β) a 1, por consiguiente, la ecuación queda representada como: 1 - β. En distintas disciplinas, el valor aceptado de la probabilidad de error tipo I o α equivale, convencionalmente, al nivel 0,05 o 0,01. Contrariamente, el nivel de la probabilidad de error tipo II o β no se halla unánimemente acordado, por lo que en algunas corrientes se asume el mismo valor de α. Desde este punto de vista, la potencia exigida debería ser ≥ 0,95 a nivel 0,05 (1 - 5). No obstante, algunos especialistas –entre los que se encuentra el pionero en este tipo de cálculos, a saber: Cohen (1977, 1988) – recomiendan considerar 0,20 como el cómputo estándar de β; así pues, el poder estadístico quedaría fijado en ≥ 0,80 (Erdfelder, Faul y Buchner, 1996: 2).

La importancia de tomar en cuenta la potencia estadística en el diseño del corpus de estudios radica en que existe una relación entre este índice y el tamaño muestral. En consecuencia, el N debería estar cimentado sobre la magnitud estadística que se alcanzaría con él (Erdfelder *et al.* 1996; Faul *et al.*, 2007, 2009; Cárdenas y Arancibia, 2014). Conforme con esto, algunos lingüistas han optado por tomar en consideración el poder con el fin de bosquejar los corpus; pero aún siguen siendo poca las investigaciones sobre el lenguaje en las que se definan las muestras a partir de este cálculo (cf. Quesada, 2007, Aravena y Hugo, 2016, Aravena y Quiroga, 2018). En contra de esta tendencia, el número de informantes de esta tesis doctoral se precisó en función de este requisito metodológico, para lo que se recurrió al programa G*Power 3.

G*Power es un programa computacional de acceso libre y gratuito, cuya función principal es calcular la potencia para muchas pruebas estadísticas de uso común en las ciencias sociales (Faul, Erdfelder, Buchner y Lang, 2009: 1149). Literalmente, "G*Power is a tool to compute statistical power analyses for many different *t* tests, *F* tests, $\chi 2$ tests, *z* tests and some exact tests. G*Power can also be used to compute effect sizes and to display graphically the results of power analyses" (Heinrich-Heine-Universität Düsseldorf, 2023). Fue creado por un equipo de investigadores de la Heinrich-Heine-Universität Düsseldorf integrado por Faul, Erdfelder, Buchner y Lang. Una de las primeras versiones de esta aplicación se dio a conocer por estos, en 1996. En 2007, Faul, Erdfelder, Lang y Buchner presentaron una nueva propuesta denominada G*Power 3, la cual se actualizó en 2009.

En consonancia con la exposición previa, G*Power 3 permite –por un lado– proyectar el número de informantes requerido para que los resultados de los análisis estadísticos logren una magnitud efectiva ≥ 0,8. Para esto se realiza un cálculo *a priori* o antes de iniciar el estudio, en el que convergen las cifras del nivel de α, el tamaño del efecto (*d*) y el valor de la potencia deseada (1-β).

Por otro lado, el software posibilita, justamente, conocer el índice del poder que posee una muestra ya construida. En este caso, se lleva a cabo un análisis *post hoc* o con el trabajo de campo ya hecho, para lo cual se necesita el total de informantes, el nivel de α y *d* solicitado (Erdfelder *et al.*, 2009; Quesada, 2007; Cárdenas y Arancibia, 2014). A continuación, se detallan estas unidades analíticas:

- **El n muestral o cantidad de datos**, se refiere al número de participantes o cantidad de unidades lingüísticas con las que se cuenta empíricamente para el desarrollo del estudio.
- **La probabilidad de error Tipo I o α**, corresponde al intervalo de confianza. En las ciencias sociales –y, por ende, en Lingüística– equivale convencionalmente al nivel de 0,05. En otras palabras, la confianza de los datos corresponde a 95 %, mientras que la probabilidad de error tipo I es 5 %.
- **Tamaño de efecto (d)**, se refiere a la magnitud de las diferencias observables de los promedios de los grupos en interacción. En otras palabras, es "una medida de cuán profunda o fructífera fue la intervención, es decir, cuál es la magnitud del efecto del tratamiento" (Quesada, 2007). El tamaño del efecto puede ser *bajo/débil*; o *medio/moderado*; o *alto/fuerte*.

Debe destacarse que los cómputos de *d* son convencionales y dependen del tipo de estadístico utilizado en los análisis de los datos. A manera de ilustración, un tamaño de efecto débil o bajo corresponderá a un valor igual a 0,20 cuando se aplique la prueba Anova de una vía (ANOVA: fixed effects, omnibus, one-way). Por el contrario, *d* equivaldrá a 0,80 cuando el contraste sea fuerte o alto, con el mismo test.

En este mismo orden, el volumen muestral requerido dependerá de la intensidad decidida para los análisis. Por ende, si la magnitud de las diferencias de las medias es alta, el N será menor; empero si *d* es bajo o débil, el número de datos o informantes aumentará. En la tabla 3, a continuación, se exhiben algunas tasas originadas a partir de la relación entre el poder y el tipo de estadístico para la determinación del tamaño muestral.

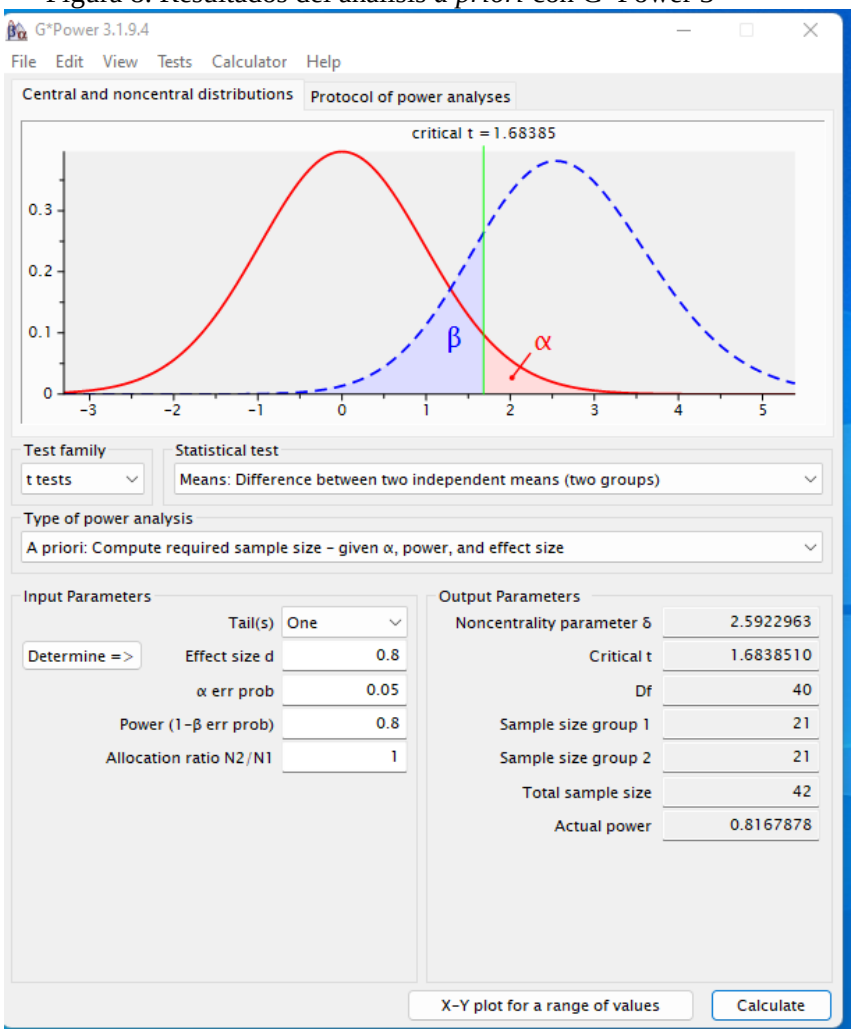
Tabla 3. Valores del tamaño de efecto para pruebas t y Anova con G*Power 3

Pruebas	*d* bajo	N muestral	*d* medio	N muestral	*d* alto	N muestral
t	0,20	620	0,50	102	0,80	42
F	0,10	969	0,25	159	0,40	66

Entonces, con base en estos planteamientos, puede deducirse que el número de participantes necesarios en una investigación, que permita establecer resultados significativos (Mafokosi, 2009), dependerá del tamaño del efecto. En los siguientes párrafos, se explica con más detalles cómo determinar el N muestral a partir de los cálculos de potencia estadística.

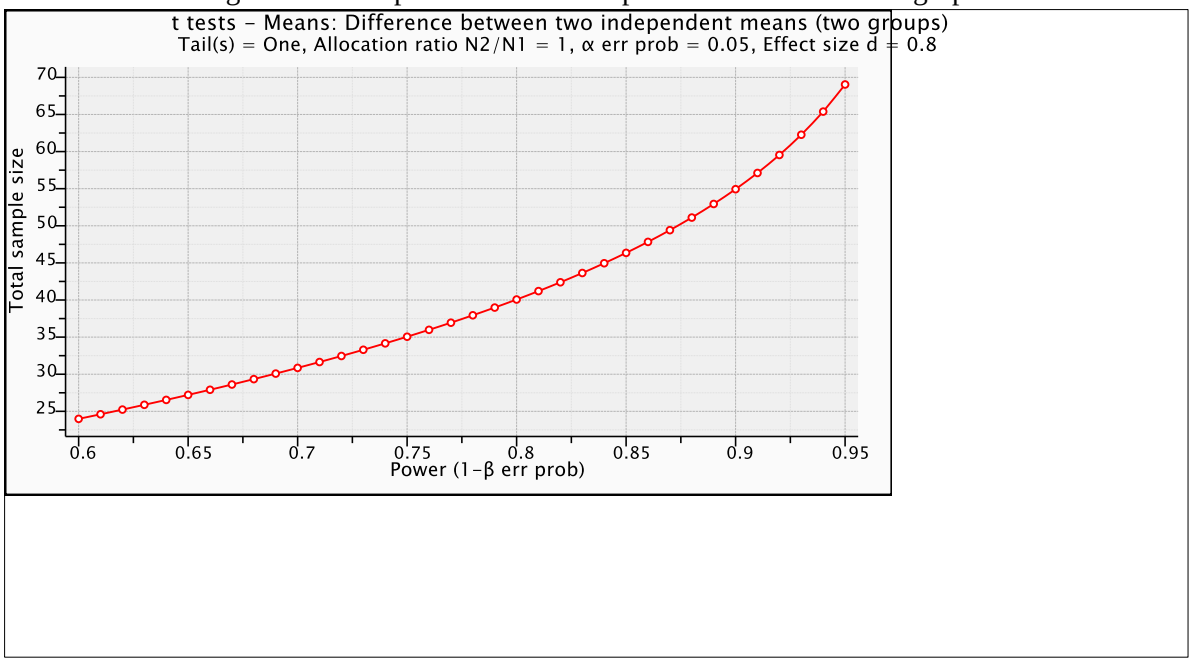
Primero, si se desea precisar el número de colaboradores (N) para construir un corpus, con alcance de potencia estadística ≥ 0,80; un α = 0,05 y tamaño de efecto alto (*d* = 0,80); en el que los datos se organizan en dos grupos independientes –por lo que los análisis se basarían en los cálculos de t de Student–, debe realizarse un análisis *a priori* con G*Power. Con los datos suministrados, la aplicación arroja que el N debe ser igual a 42 sujetos, 21 por cada grupo, como puede observarse en la Figura 8.

Figura 8. Resultados del análisis *a priori* con G*Power 3



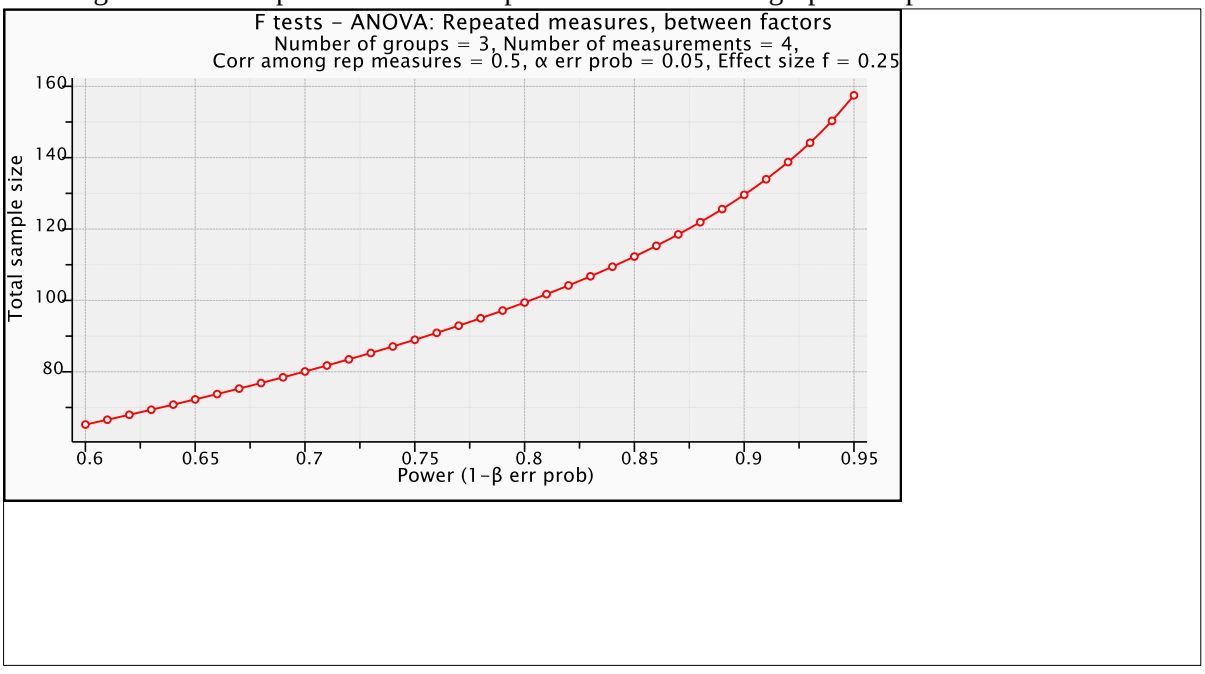
Igualmente, el programa ofrece un gráfico (X – Y *plot*) en el que se observan los grados de potencia que pueden alcanzarse con diferentes N muestrales, a partir de los datos que se le hayan proporcionado al software, como se ilustra en la Figura 9.

Figura 9. X – Y plot de G*Power 3 para una muestra de dos grupos



Segundo, análisis *a priori* para determinar el N requerido para alcanzar una potencia estadística ≥ 0,8, en una muestra compuesta por tres grupos independientes con un efecto moderado (d = 0,25) válido para los test de la familia F y un α = 0,05. De acuerdo con la Figura 10, el corpus debería estar constituido por, al menos, 102 sujetos repartidos equitativamente (34 participantes) en los tres grupos. Por el contrario, se necesitarían 132 para llegar a una potencia igual a 0,9.

Figura 10. X – Y plot de G*Power 3 para una muestra de 3 grupos con potencia moderada.

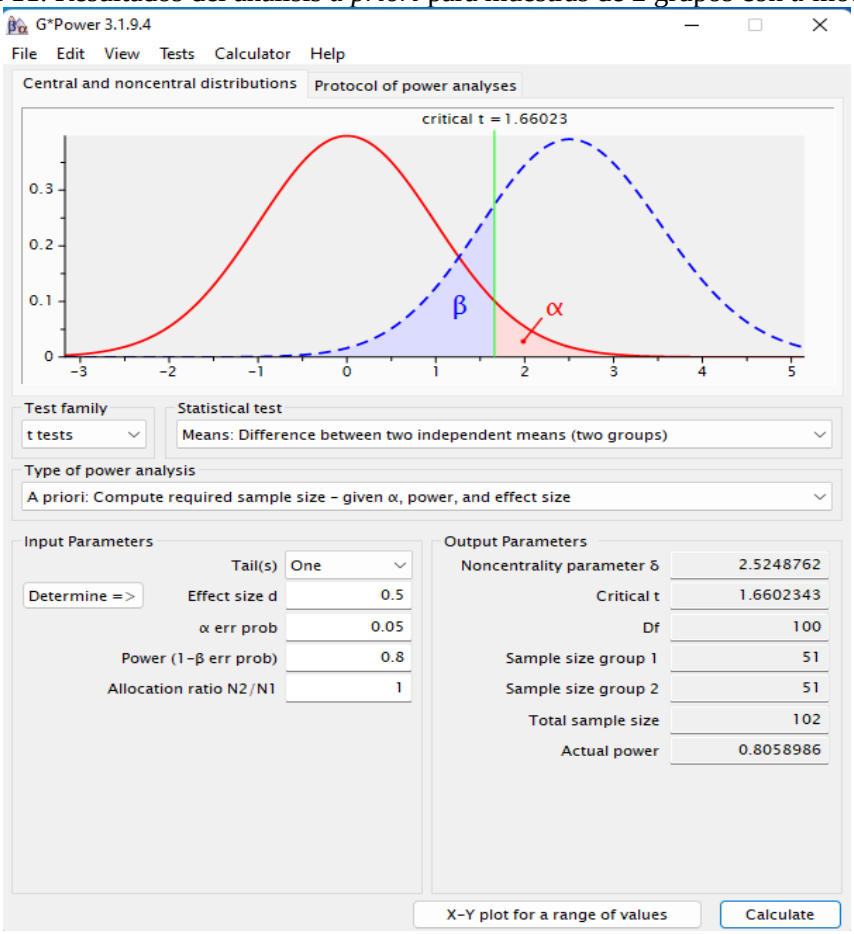


En consonancia con los fundamentos expuestos por Quesada (2007) y Cárdenas y Arancibia (2014), en esta tesis doctoral se optó por definir el tamaño de la muestra y el umbral de significación, a partir del criterio de potencia estadística calculado a través de G*Power 3 (Erdfelder *et al.*, 2007, 2009). En atención a lo cual, en el análisis *a priori* se tomaron en cuenta los siguientes valores:

- α = 0,05
- d = medio o moderado
- potencia estadística (1 - β) = 0,8

A fuer de que los datos estarían distribuidos en dos grupos según las variables de estudio, ya sea por *sexo* (hombres y mujeres), o por *formato de prueba* (papel y digital), o por *carrera* (Educación Básica y Letras Hispánicas), o ya sea por nivel del curso (1.er y 4.o Año), a los que se les aplicaría la prueba t de Student, el cálculo *a priori* con G*Power 3 arrojó que el N muestral debía ser igual a 102 individuos (51 por cada conjunto), de manera que la potencia estadística fuese ≥ 0,8. En consecuencia, para esta tesis, la meta era contar con un corpus que satisficiera estos criterios cuantitativos, como se aprecia en la Figura 11.

Figura 11. Resultados del análisis *a priori* para muestras de 2 grupos con *d* moderado

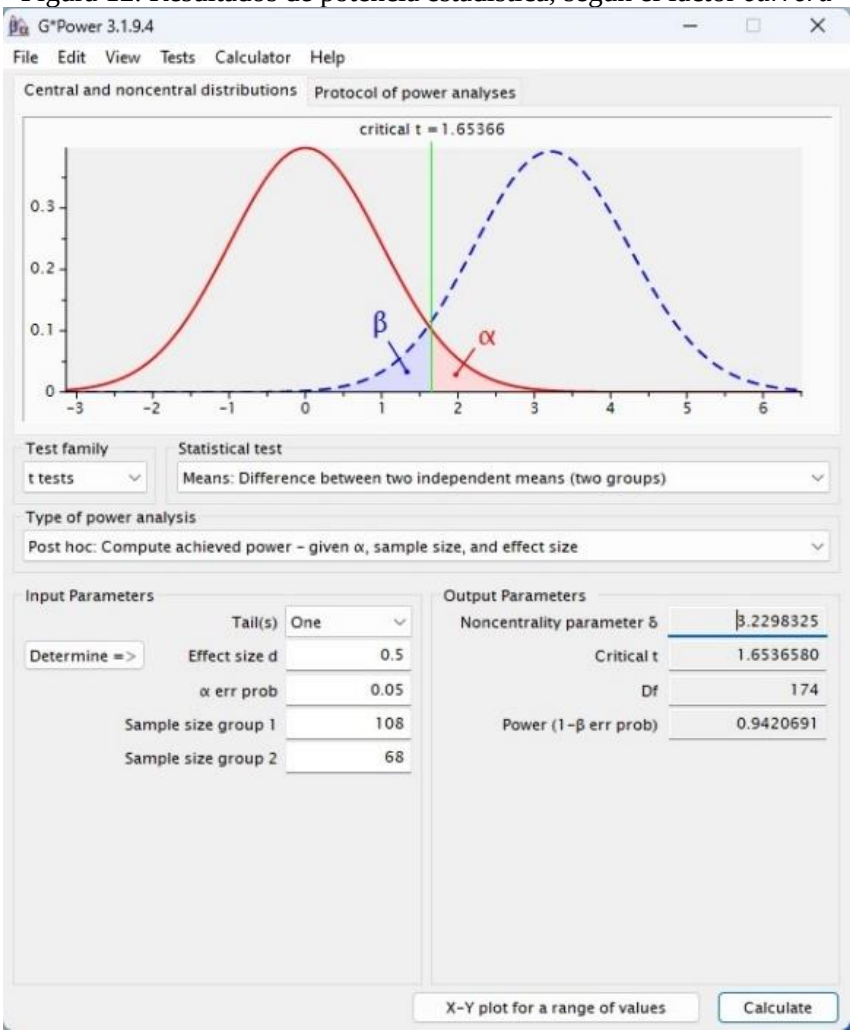


Una vez que se llevaron a cabo tanto los trabajos de campo para la recolección de los datos, como la depuración de los listados de palabras, el N muestral final de esta tesis quedó definido por las respuestas de 264 encuestados. Es decir, un cómputo por encima del estimado en el diseño metodológico *a priori*. Estos datos se distribuyen en tres grupos, a saber:

- grupo 1, compuesto por las respuestas de 108 discentes de la carrera de Educación Básica;
- grupo 2, integrado por las encuestas de 68 estudiantes de Letras Hispánicas; y
- grupo 3, conformado por las listas de palabras de 88 alumnos de Letras Hispánicas, quienes realizaron los test en formato digital.

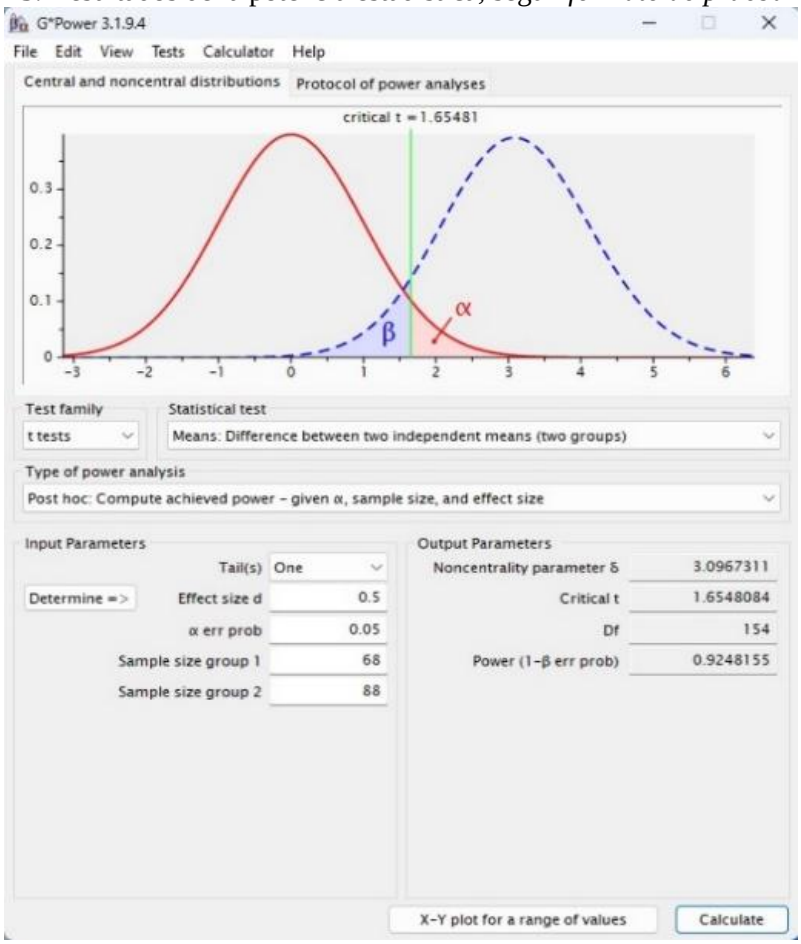
Con estas cifras, los análisis *post hoc* de potencia estadística realizados con G*Power 3 arrojaron los siguientes resultados: i) en la comparación de los datos respecto al factor *carrera* –en la que se tomaron en cuenta únicamente las respuestas elaboradas en formato papel–, la potencia estadística equivale a 0,94 > 0,80, según se observa en la Figura 12, a continuación.

Figura 12. Resultados de potencia estadística, según el factor *carrera*



Por su parte, en el contraste de los dos grupos definidos a razón de la variable *formato de las pruebas* –en el que se consideraron solamente las respuestas de los alumnos de Letras Hispánicas–, el índice alcanzó un cómputo igual a 0,92 > 0,80, como se aprecia en la Figura 13. En síntesis, los resultados de ambas comparaciones muestrales se traducen en que la distribución de los sujetos es estadísticamente confiable para los análisis léxicos basados en la familia t-test, para las variables *carrera* y *formato de prueba*.

Figura 13. Resultados de la potencia estadística, según *formato de prueba*



2.3.2. Tamaño de la muestra: porcentaje de matrícula

Con la finalidad de precisar aún más la fiabilidad de la distribución de los datos del corpus, se revisaron los padrones de las matrículas de Educación Básica y Letras Hispánicas de la PUC publicados por el Concejo Nacional de Educación (CNED). La CNED es un organismo del Estado chileno cuya misión es "cautelar y promover la calidad de la educación parvularia, básica, media y terciaria en el marco de los sistemas de aseguramiento de la calidad de la educación escolar y superior" (CNED, 2022). Aunado a esto, en el caso de la Facultad de Letras, los índices facilitados por el CNED

se corroboraron con los catastros proporcionados por la dirección académica del programa universitarios antes mencionado.

En las tablas 4 y 5, se exponen los porcentajes que representa cada N muestral en relación con el total de la población teóricamente elegible, la cual se basa en los indicadores proporcionados por las instituciones. En dichas tablas pueden apreciarse las siguientes cifras por carrera: i) Vacantes, se refiere a la cantidad de alumnos que pueden recibirse cada año; ii) Matrícula, concierne al total de estudiantes regulares inscritos en el programa. Este contempla a todo el alumnado, desde el primer al último año de formación; es decir, la suma de todas las cohortes. iii) Población elegible, implica al número aproximado de estudiantes de 1.er y 4.o año; quienes serían los candidatos específicos del estudio y, por ende, se les podría aplicárseles los cuestionarios de disponibilidad léxica. iv) Participantes, compete justamente al número de sujetos que respondió efectivamente los test de DL. v) % I, corresponde al porcentaje que representan los participantes en relación con la población elegible. Por último, vi) % II, muestra el porcentaje de los voluntarios respecto a la matrícula.

Tabla 4. Características de las muestras en función de las matrículas del año 2022

Carrera	**Vacantes**	**Matrícula**	**Población elegible**	**Participantes**	**% I**	**% II**
Educación Básica	150	621	194	**108**	55,67	17,39
Letras Hispánicas	60	264	135	**68**	50,37	25,75
Total	210	885	329	**176**	53,49	19,88

Como puede apreciarse en la tabla 5, los porcentajes de los individuos por carrera llegan al 50 %, siendo el mayor valor el del alumnado de Educación Básica, que alcanzó un 55,67 %; mientras que los de Letras Hispánicas son el 50,37 % de la población accesible. No obstante, estos valores distan mucho cuando se contabiliza la matrícula total de cada carrera. En este caso, los grupos tienen porcentajes por debajo del 30 %, Letras Hispánicas = 25,75 %, y Educación Básica = 17,39 %.

En función de las respuestas recibidas gracias a la plataforma digital, se decidió tratar únicamente las listas de palabras de un grupo de Letras Hispánicas; puesto que se recibieron pocos cuestionarios del alumnado de Educación Básica (menos de 20 voluntarios). En definitiva, se contó con 88 encuestas totalmente respondidas por los colaboradores de LH, quienes guardaban los requisitos mínimos exigidos para el estudio. En la tabla 6, seguidamente, se muestran los números.

Tabla 5. Definición de la muestra digital, en función la matrícula del 2022

Carrera	**Vacantes**	**Matrícula**	**Población elegible**	**Participantes**	**% I**	**% II**
Letras Hispánicas	60	263	136	88	64,70	25,85

Sobre la tabla 6, puede sostenerse que los porcentajes muestran que los datos representan el 64,70 % de la población accesible; empero, un 25,85 % del total de la matrícula del 2022. En este punto resulta importante señalar que los porcentajes de los participantes en relación con el catastro de

la escuela de Letras Hispánica son teóricos o aproximados, porque la muestra se constituyó entre el segundo semestre de 2020 y el primer semestre de 2022.

2.4. Informantes

Una vez explicado el tamaño de la muestra, en este epígrafe se describen las características generales de los participantes, según los factores: i) *carrera* y ii) *formato de pruebas*. Además, se indican cuáles fueron los mecanismos de acceso a la población.

Al recibir la aprobación del comité de ética, se dio marcha blanca a la solicitud de los permisos pertinentes para la aplicación de las pruebas a estudiantes de Educación Básica y Letras Hispánicas. En primera instancia, se realizaron las reuniones –vía Zoom– con las directoras de carreras, jefas de departamentos y coordinadoras de investigación, con el objetivo de brindarles toda la información concerniente al proyecto de tesis doctoral. Específicamente, se les explicó cómo sería la captación y aplicación de las pruebas. En segunda instancia, ya con los permisos obtenidos, se contactó a los docentes que dictaban asignaturas de 1.er y 4.o año de las respectivas carreras, quienes a su vez permitieron:

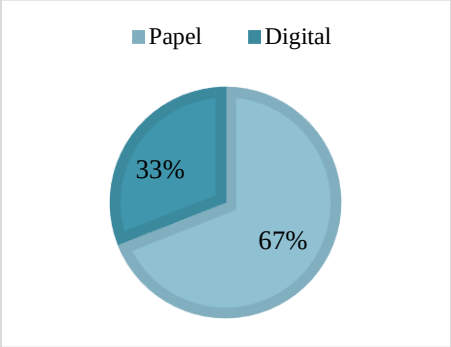
1) Respecto al método tradicional de toma de datos en formato papel, asistir a sus clases un día y hora previamente acordado. Una vez en el aula, se les indicó a los alumnos cuáles eran los objetivos del estudio, los compromisos éticos y de integridad en la investigación (la realización de las pruebas es voluntaria; los datos son anónimos y confidenciales), y el método de aplicación de las encuestas. Así pues, los test los contestaron los sujetos que dieron su consentimiento informado.
2) En cuanto al método alternativo de recogida de datos (soporte digital), vía Zoom durante el periodo de clases *online* por la pandemia, a los estudiantes conectados al aula virtual se les brindó toda la información del proyecto. Luego, se les envío el vínculo electrónico de la encuesta a quienes voluntariamente aceptaron participar. Debe señalarse que los colaboradores leyeron la carta de consentimiento informado antes de empezar las pruebas. Debe recalcarse que los datos son anónimos y confidenciales.

La recolección de todos los datos se realizó entre el segundo semestre de 2020 y el primer semestre del 2022. El alargamiento del tiempo de este proceso se debió a los problemas externos que surgieron durante el desarrollo de la investigación, específicamente: el estallido social y las cuarentenas por la pandemia de la covid-19. Estos eventos impidieron y limitaron el acceso a la población elegible. A pesar de los inconvenientes, se recogieron 264 test de disponibilidad léxica, completamente válidos, mediante ambos métodos, de sujetos cuyas características generales eran:

1) estudiantes universitarios regulares de las carreras de Educación Básica y Letras Hispánicas, de la Pontificia Universidad Católica de Chile; institución acreditada por la Comisión Nacional de Acreditación de Chile (CNA-Chile);
2) cursantes del 1.er o 4.o año de ambos programas universitarios.

De las 264 encuestas válidas, 108 voluntarios son cursantes de Educación Básica, lo que representa el 41 %, mientras que el restante 156 eran alumnos de Letras Hispánicas, un 59 %. Sin embargo, del N total, 176 sujetos respondieron las encuestas por medio del método tradicional en formato papel, lo que se traduce en un 67 %. Por su parte, 88 cuestionarios fueron contestados a través del método alternativo en formato digital, lo que equivale al 33 %, según se ilustran a continuación.

Gráfico 3. Porcentajes de los participantes, según *formato de pruebas*



Más detalladamente, en relación con los factores *Carrera* y *Formato de las pruebas*, las respuestas léxicas de los 264 participantes se organizaron en tres muestras, como se explica a continuación: i) la primera está constituida por 108 cuadernillos en formato papel de estudiantes de Educación Básica, lo que representa un 40,90 %; ii) la segunda está compuesta por las listas de palabras, también en papel, de 68 alumnos de Letras Hispánicas, un 25,75 %; por último, iii) la tercera muestras está integrada por las respuestas, en soporte digital, de 88 discentes de Letras Hispánicas, lo que se traduce en un 33,33 % del N. En la Tabla 6 se detalla la distribución de los informantes.

Tabla 6. Distribución de los informantes por muestra

Muestras	**n**	**%**
Muestra 1: Ed. Básica en papel	108	40,90
Muestra 2: Letras H. en papel	68	25,75
Muestra 3: Letras H. digital	88	33,33
N total	264	
Potencia estadística (1-β) = 0,9974537		

En consonancia con las finalidades de la investigación, las muestras se reagruparon en dos conjuntos. El primero, al que se ha denominado subcorpus 1, está conformado por las 176 listas de palabras que fueron elaboradas en los cuadernillos de papel por los cursantes de Educación Básica y

Letras Hispánicas. El objetivo de esta combinación, en la que se suman los materiales de las muestras 1 y 2, es contrastar el léxico disponible de los estudiantes de la Facultad de Educación con el de los de la Facultad de Letras, de manera que se pudiera determinar si el factor *carrera* tiene alguna incidencia en el vocabulario de los encuestados, respecto a los ocho centros de interés planteados. En la tabla 7 de aprecia esta muestra.

Tabla 7. Organización de los encuestados por carrera en el subcorpus 1

Subcorpus 1			
Muestra 1: Ed. Básica		Muestra 2: Letras H.	
n	%	n	%
108	61,36	68	38,64
176			
$1 - \beta = 0{,}9420691$			

Por su lado, el segundo conjunto, llamado subcorpus 2, está integrado únicamente por las respuestas léxicas de los 156 encuestados de Letras Hispánicas, quienes respondieron las pruebas de disponibilidad léxica a través de cuadernillos en dos soportes distintos, a saber: en papel, al tenor del método tradicional (cf. Gougenheim *et al.*, 1964; Dispolex, 2023), y en digital, que es el método alternativo propuesto en esta tesis. Específicamente, el subcorpus 2, entonces, lo componen las listas de palabras de las muestras 2 y 3, como se detalla en la tabla 8, seguidamente. Esta organización muestral tenía el objetivo de evaluar el caudal léxico recolectado de manera remota mediante la plataforma digital creada para este estudio. Respecto a los instrumentos aplicados para tomar los datos, se describen en el subapartado siguiente.

Tabla 8. Organización de participantes en el subcorpus 2

Subcorpus 2: Letras Hispánicas			
Muestra 2: en papel		Muestra 3: digital	
n	%	n	%
68	43,58	88	56,42
156			
$1 - \beta = 0{,}9248155$			

2.5. Instrumento de recolección de datos

Una vez que se ha conocido la conformación del corpus, en este subapartado se describen y explican los tres instrumentos de recolección de datos, a saber: el cuestionario sociológico, los test de disponibilidad léxica y la encuesta de prácticas lectoras.

2.5.1. Cuestionario sociológico

En el siglo XX hubo un cambio de paradigma en los estudios de la lengua, puesto que empezaron a surgir nuevas maneras de abordar el lenguaje, en las que se consideraban aspectos sociales, funcionales y cognitivos, en los que no se enfocaban las corrientes dominantes de la época: el estructuralismo y el generativismo. Así pues, entraron a la esfera de la Lingüística las corrientes de corte funcional y cognitiva. En este nuevo contexto, nace la sociolingüística, disciplina que busca determinar la relación entre los hechos lingüísticos y los sociales (López Morales, 2015; Moreno Fernández, 2017; Silva-Corvalán y Enrique-Arias, 2017; Blas Arroyo, 2019, entre otros). Al respecto, Cezario y Votre (2013: 141) señalan que: "La sociolingüística es un campo que estudia la lengua en su uso real, teniendo en cuenta las relaciones entre la estructura lingüística y los aspectos sociales y culturales de la producción lingüística"[9]. En consideración a lo anterior, se considera la lengua como una institución social, por lo que en su estructuración y función existen marcas sociales de distintas índoles. En consecuencia, la sociolingüística pone en el núcleo de sus estudios aquellos fenómenos del lenguaje que los modelos formalistas consideraban marginales (Almeida, 2003).

Las investigaciones sociolingüísticas iniciaron con el nivel fonético-fonológico, en el que se alcanzaron resultados interesantes sobre variación y cambio lingüístico, que fueron cimentando su aparataje teórico y metodológico. La extensión de los postulados de esta nobel disciplina a los demás niveles de la lengua no fue tan sencillo de justificar; pese a esto, se han desarrollado estudios exhaustivos y sólidos acerca de variación y cambio lingüístico en fenómenos morfológicos, sintácticos y pragmáticos (cf. Silva-Corvalán y Enríquez-Arias, 2017).

En el nivel léxico, la sociolingüística enfrentó los inconvenientes teóricos y metodológicos observados en los demás, entre los que pueden señalarse: la discusión sobre la sinonimia absoluta (López Morales, 2015: 105) y el uso de encuestas en vez de entrevistas (Almeida, 2023: 75). De manera específica, la discriminación de las unidades de análisis fue más compleja. Sin embargo, se han llevado a cabo estudios de variación léxica significativos –al respecto, pueden consultarse los trabajos enmarcados en el proyecto Varilex (2023), que busca *grosso modo* "conocer la situación actual del léxico español del mundo", coordinado por *el profesor* Hiroto Ueda, de la Universidad de Tokio–.

En este panorama, la disponibilidad léxica resulto importante para la sociolingüística, puesto que la metodología aportada por la DL ha permitido ahondar en la descripción y explicación de la

[9] Nuestra traducción de: "A sociolinguística é uma área que estuda a língua em seu uso real, levando em consideração as relações entre a estrutura linguística e os aspectos sociais e culturais da produção linguística" (Cezario y Votre, 2013: 141)

estratificación social del léxico de las comunidades de habla. De hecho, ha existido una vinculación interdisciplinar entre ambas disciplinas (López-Morales, 1995-1996; Ávila Muñoz y Villena, 2010; Moreno Fernández, 2017, entre otros). El arqueo bibliográfico evidencia los análisis en los que se correlacionan el caudal léxico y factores sociales, como: sexo, edad, zona de residencia, entre otros. Si bien, el objetivo de esta tesis doctoral no es explorar la variación *per se* en el léxico disponible de los alumnos de Educación Básica y Letras Hispánicas, sí interesa observar cómo las características estadísticas del caudal léxico de los encuestados pueden estar motivadas por aspectos sociológicos y socioeducativos.

En consonancia con lo antes expuesto, este trabajo cuenta con una *encuesta sociológica*, que es un aparato analítico centrado en la recolección de la información diastrática pertinente, con el fin de estratificar sociolingüísticamente a los participantes de este estudio. Debe resaltarse que, en el diseño del cuestionario, se tomaron en cuenta las dimensiones éticas de la investigación científica, por lo que los datos son anonimato y confidenciales.

La encuesta sociológica tiene en su encabezado un código numérico que se ha rotulado para mantener ordenados los materiales de un mismo voluntario y, además, para cuidar la identidad de los sujetos, a través de la anonimidad. Igualmente, aparece un anunciado que recuerda las disposiciones éticas de la investigación. Las preguntas del cuestionario se distribuyen en tres partes. En la primera (*Datos personales*), aparecen 7 preguntas relacionadas con categorías sociológicas generales, como: sexo, edad, lengua materna. En la segunda parte, *Datos de los padres*, se recogen los rasgos educativos, profesionales y ocupacionales de los padres del encuestado, a través de 11 preguntas. Por último, *Información educativa del encuestado*, se recopilan los datos escolares mediante 10 preguntas, como: tipo de institución donde realizó estudios de secundaria, carrera que cursa, año de la carrera, financiamiento, estudios previos, entre otras. En el caso del método digital, el número y orden de las preguntas se mantuvieron iguales que en el formato físico o en papel.

Con base en la data sociodemográfica recogida, se definieron las variables socioeducativas de esta tesis, a saber: *Sexo*, *Carrera* y *Nivel del curso*. Con la descripción de estos factores se codificaron los datos para los análisis cuantitativos y cualitativos mediante los programas computacionales Dispogen, Dispografo y SPSS. En el anexo 1, se aprecia este instrumento.

A continuación, se describe el segundo instrumento de recolección de datos.

2.5.2. Prueba de disponibilidad léxica

El segundo instrumento concierne a las pruebas de disponibilidad léxica, que son las más importante de este estudio, puesto que con ellas se recoger el léxico disponible de los sujetos referido a un centro de interés particular. En palabras de Hernández Muñoz y Tomé (2017: 100) el test de DL es una "[…] pruebas de fluidez categorial basadas en modelos instruccionales del tipo «dime/escribe todas las palabras que conozcas de la categoría Alimentos»".

En este trabajo, los test de DL –siguiendo el modelo tradicional– están conformados por cuadernillos de ocho hojas de papel blancas, tamaño carta, una por cada área nocional analizada. En la parte superior de cada prueba se encuentra rotulado el código de identificación del participante. Igualmente, se lee la sigla CI (centro de interés) más el número y la sigla correspondiente del actualizador, como se ilustra seguidamente. El código *CI1: LL* significa *centro de interés 1: la lectura*. En la siguiente hoja estaba, entonces, *CI2: EP*, a saber: *centro de interés 2: el profesor*, y así sucesivamente hasta el último eje temático. Esta nomenclatura fue ideada con la finalidad de sistematizar los materiales a través de la ordenación e identificación de las áreas nocionales.

En las instrucciones para la realización de las pruebas de DL en formato papel (método tradicional) se les indicaba a los voluntarios que debían escribir todas las palabras que se les vinieran a la cabeza sobre el tópico del centro de interés enunciado. También se les hizo la acotación de que no debían reportar conceptos ni definiciones ni descripciones. Además, se les informó que no había palabras "buenas" ni "malas"; "correctas" ni "incorrectas".

Las ocho áreas nocionales de este estudio contemplan centros de interés tradicionales –es decir, aquellos planteados en los trabajos franceses e integrados en el PPHLD–, a saber: *la escuela: muebles y materiales*, *juegos y distracciones*, *partes del cuerpo* y *comidas y bebidas*. También se preguntó acerca de un actualizador utilizado en estudios recientes, concretamente: *la educación*[10]. Por último, se han propuesto tres ejes temáticos nuevos: *la lectura*, *el profesor* y *habilidades y cualidades docentes*. El orden de presentación y análisis de estos campos nocionales es:

1. La lectura
2. El profesor
3. La educación
4. Juegos y distracciones
5. La escuela: muebles y materiales

[10] En el caso de este centro de interés, debe acotarse que el nombre difiere del de Herranz (2020) tan solo por la eliminación del artículo.

6. Habilidades y cualidades docentes
7. Partes del cuerpo
8. Comidas y bebidas

Respecto a las respuestas, están eran listas abiertas de palabra, por lo que no hubo un número estipulado de lexías por centro de interés, para lo que los sujetos contaban con dos minutos por área nocional para elaborar sus listados. Esto significa que la prueba de DL duraba 16 minutos.

En cuanto a los cuestionarios de disponibilidad léxica en formato digital, estos mantuvieron a grandes rasgos todos los aspectos metodológicos del formato en papel, con las salvedades derivadas del tipo de soporte. En este marco, las instrucciones se transliteraron –mediante un lenguaje claro y con ejemplo–, para que pudieran ser leídas por los participantes. Estas se exponían en una ventana autónoma que se mostraba previamente a la aparición de las hojas de respuestas de cada centro de interés. Las cajas electrónicas para la recolección de las listas de palabras tenían visible únicamente el comando y el nombre del área nocional. Las ventanas digitales de respuesta cada CI –al igual que en el modelo tradicional– permanecían abiertas solamente por dos minutos. De manera que, una vez acabado ese tiempo, se cerraba automáticamente la prueba y, segundo después, se abría el test del siguiente eje temático.

Al terminar el tiempo de respuesta del último centro de interés, *comidas y bebidas*, se habilitaba –tanto en método tradicional como en el alternativo– el último instrumento: la encuesta de práctica lectoras. Este se describe en el siguiente epígrafe.

2.5.3. Encuesta de prácticas lectoras

La lectura es una actividad cognitiva que integra procesos de reconocimiento de signos, comprensión, interpretación, entre otros. Además, es una de las maneras más extendida de construir y generar conocimiento. Como herramienta en la enseñanza formal, la lectura se va complejizando a medida que se avanza en el sistema educativo: desde la lectura corta y lineal de los primeros años, cuyo objetivo es aprender a leer; pasando por el aumento de textos y actividades asociadas a ella, con el foco puesto en leer para aprender; hasta llegar a mayores exigencias como la comprensión, relación y posicionamiento crítico (Muñoz *et al.*, 2012). En la vida universitaria, la lectura se ve motivada por la intención que tienen los estudiantes por adquirir nuevos conocimientos, profundizar en ellos, tener la capacidad crítica de confrontar los postulados aprendidos y desarrollar habilidades profesionales. Sin embargo, llegar a estos puntos de abstracción y relación con la lectura en la universidad requiere

–además del hábito, compromiso y motivación– que el alumno tenga conocimientos previos acerca de la materia y el léxico, tanto el general como técnico-especializado (Muñoz *et al.*, 2012).

El arqueo bibliográfico muestra un alto número de investigaciones sobre la lectura, particularmente desde el ámbito pedagógico (Larrañaga *et al.*, 2008; Manresa, 2009; Munita, 2014; Parrado *et al.*, 2018; Álvarez-Álvarez y Diego-Mantecón, 2019, entre otros). Por su parte, léxico-estadísticamente han sido poco abordadas las relaciones entre el caudal léxico de estudiantes universitarios y sus prácticas lectoras, aunque se está observando un progresivo interés por ahondar acerca de este tema desde la disponibilidad léxica (cf. Álvarez-Álvarez y Diego-Mantecón, 2019; Santos Díaz, 2017a, 2020; Juárez, 2019; Trigo Ibáñez *et al.*, 2019; Santos Díaz *et al.*, 2021; Santos Díaz y Juárez, 2022). Por lo que, en consideración a los objetivos de esta investigación, se ha aplicado una encuesta de prácticas lectoras, con el fin de recabar información acerca de las prácticas lectoras de los participantes, de manera que pudiera analizarse la incidencia de los factores relacionados con la lectura sobre el LD.

Entonces, después del test de disponibilidad léxica, se aplicó el cuestionario de prácticas lectoras, que estaba compuesto por 13 ítems, organizados en tres partes. En la primera parte, se encontraban las preguntas relacionadas con el gusto por la lectura y el uso de internet. En la segunda parte, estaban los puntos interrogativos enfocados en la motivación por la lectura. Y, en la última parte, se solicitaba información acerca de las prácticas lectoras de textos optativos (los que no tenían que ver con los programas de estudios) y obligatorios (los propios de las carreras). Las preguntas de la encuesta de prácticas lectoras eran de varios tipos: abiertas, de selección múltiple y de escala psicométricas o escala de Likert, esto es "una escala sumatoria" (Briones, 1990: 137). Al igual que con las otras dos herramientas de recopilación de datos, la página web de esta tesis mantuvo las instrucciones y preguntas en el mismo orden y del mismo tipo: abiertas, selección múltiple y escala de Likert, de la batería de prácticas lectoras aplicada de forma tradicional.

Debe señalarse que este instrumento es una adaptación de la encuesta de hábitos lectores propuesta por el proyecto Fondecyt Regular n° 1170779, dirigido por la profesora Carla Muñoz. Además, se ha tomado en cuenta los trabajos de Ávila Muñoz (2007), Larrañaga, Yubero y Cerrillo (2008) y Santos Díaz (2017a). Para más detalles puede revisarse este cuestionario en el Anexo 2.

En el siguiente apartado, se reseñan los procedimientos por medio de los cuales se aplicaron los instrumentos descritos en este epígrafe.

2.6. Métodos de recolección de datos

Como se ha indicado en el subapartado 2.4, el corpus general está compuesto por las listas de palabras de 264 voluntarios, quienes contestaron las pruebas de DL a través de dos formatos distintos (en papel y digital), pero que cumplían *grosso modo* los mismos criterios metodológicos. En los próximos párrafos se explican las singularidades de ambos procedimientos de recolección de datos.

2.6.1. Método o técnica tradicional

Se ha denominado método o técnica tradicional al procedimiento de toma de datos presencialmente, según la propuesta de Gougenheim *et al.* (1954, 1964), la cual ha sido asumida –con las respectivas adecuaciones– por el PPHLD. Concretamente, este método comporta *grosso modo* cuatro pautas generales, las que se han tomado en cuenta en esta tesis también.

- La primera concierne a la presencialidad: las pruebas se administraron directamente a la población elegible en las aulas de clases, en los horarios previamente convenidos con los profesores de las cátedras que dieron el permiso. Se recolectaron los listados de palabras de quienes dieron su consentimiento informado.
- La segunda se refiere al modo: los encuestados escribieron a mano sus listados de unidades léxicas en cuadernillos de papel, entregados por el investigador responsable.
- La tercera atañe al tiempo de aplicación de las pruebas: si bien, el periodo de respuesta por actualizador ha oscilado entre dos y cinco minutos en las distintas investigaciones (cf. Dimitrijević, 1969; Justo, 1986; Ruiz Basto, 1987; entre otros), actualmente, se ha convenido otorgar dos minutos por campo nocional, gracias al impulso del PPHLD. Este criterio se fundamentó, entre otras razones, por los resultados del trabajo de Mena Osorio (1986). En consecuencia, en esta investigación, los voluntarios contaron con dos minutos, por centro de interés, para manufacturar cada lista de palabras. Además, el inicio y termino de esta tarea fue instruido por el responsable, quien llevaba el tiempo cronometrado reloj en mano. En virtud de que el cuadernillo constaba de ocho áreas nocionales, contestar el test de DL tomó 16 minutos.
- Finalmente, la cantidad de palabras por centro de interés. Al igual que el tiempo de respuesta, este rasgo ha variado en las diferentes investigaciones de DL. En algunos casos se ha estipulado un número preciso de piezas léxicas, mientras que en otros se conciertan listas abiertas (cf. Mena Osorio, 1986; Mateo García, 1998). Este estudio se decantó por la segunda

opción, por lo que los participantes trazaron, en dos minutos, todas las unidades léxicas sin límites que se les vinieron a la cabeza sobre el estímulo mencionado.

Aunado a estos criterios, los encuestados iban contestando uno por uno los test de DL, a medida que era indicado por el investigador responsable. Así pues, no podían regresar a los listados previos a corregirlos ni tampoco a añadir nuevas piezas léxicas. Asimismo, respecto al tiempo total requerido para aplicar las pruebas, este osciló entre 30 y 45 minutos, ya que se sumaba la duración de lectura y respuestas de los cuatro instrumentos, cuyo orden de presentación fue: 1) consentimiento informado; 2) encuesta sociológica, 3) test asociativo de disponibilidad léxica, y 4) batería de prácticas lectoras. Por último, debe resaltarse que, en el contexto pandémico en el que se desarrolló este trabajo, se administraron las pruebas de disponibilidad léxica con este método, manteniendo los protocolos de bioseguridad.

2.6.2. Método digital

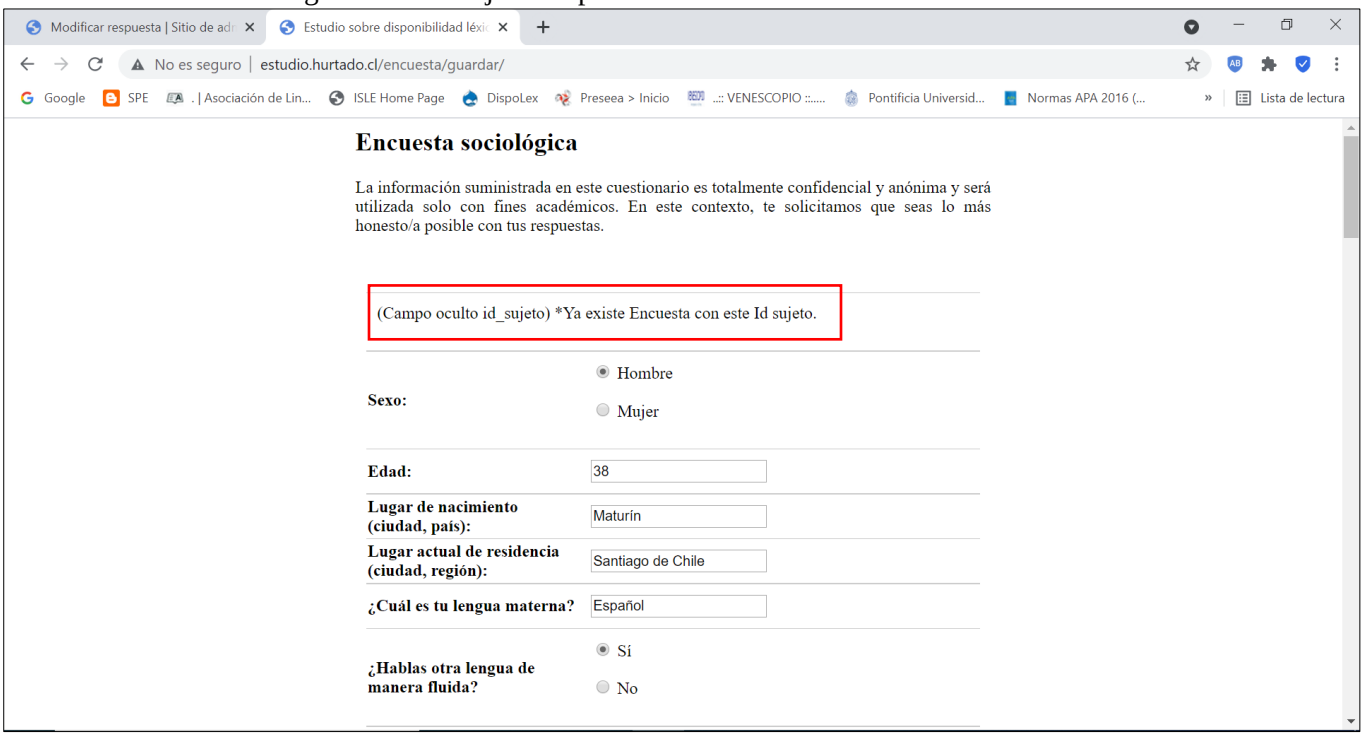
En atención a los objetivos de esta investigación, se diseñó una plataforma electrónica *ad hoc* con la que se recolectó el léxico disponible de los informantes de manera remota, a tenor de los planteamientos metodológicos de la disponibilidad léxica. La decisión de crear este instrumento digital radicó en que las plataformas de acceso abierto y gratuito, como *Google Forms* y *Monkey Survey*, que funcionan para crear encuestas, no cuenta con la opción de limitar el tiempo de respuestas, que es un criterio cardinal en este tipo de estudios. En cambio, los sitios web que sí contemplan recursos para fijar los intervalos de reacción a los test tienen acceso restringido, sus sistemas de programación suelen ser complicados y, sobre todo, no son gratis.

La página web se creó, manteniendo lo más fiel posible, los criterios metodológicos fundamentales de los trabajos de DL, en consonancia con las propuestas del PPHLD (Dispolex, 2003-2023). En este marco, las pruebas de DL de esta aplicación digital contaban con: 1) un cronometro que aseguraba los dos minutos de duración por centro de interés para que los participantes realizaran las listas de palabras. 2) Las cajas electrónicas de respuestas permitían que se escribiera un número ilimitado de piezas léxicas, mientras estuvieran abierta. Por último, 3) una vez terminados los dos minutos, las hojas digitales se cerraban completamente, sin permitir la posibilidad de retomar la terea. Inmediatamente, se abría el cajón del subsiguiente eje temático.

El acceso al sitio web es restringido. Solo pueden entrar quienes han expresado su deseo de colaborar libremente en el estudio y, por ende, han recibido, vía correo electrónico, la carta de consentimiento informado y un vínculo computacional para el ingreso, como este:

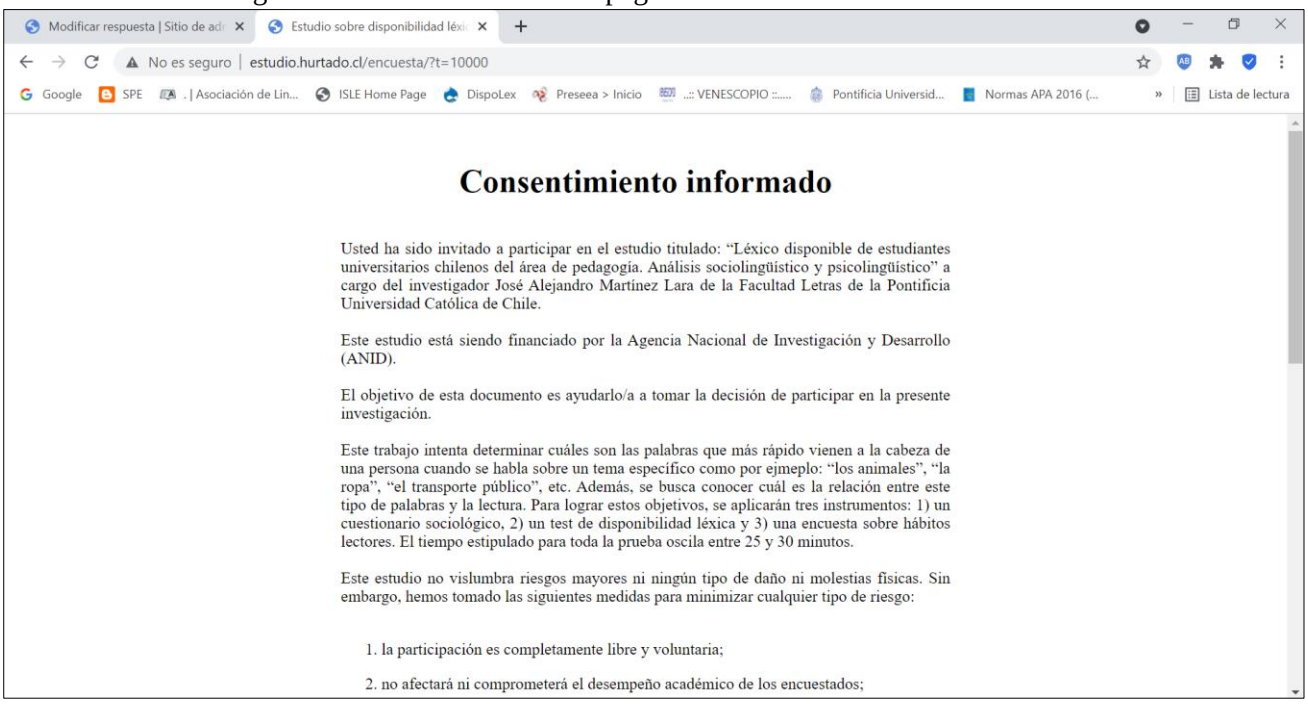
http://estudio.hurtado.cl/encuesta/?t=10000. Estos enlaces son administrados y suministrados solamente por el investigador responsable. Cada *link* es único, pero no guardan datos sensibles que pudieran identificar a los participantes. Asimismo, cada vínculo electrónico se activa una única vez. Es decir, tras haberse contestado y cerrado toda la prueba, el *link* caduca y se desecha. Esto impide el ingreso recurrente a la plataforma, evitando que los sujetos alteren sus respuestas. En caso de volver a meterse a la aplicación con una misma dirección electrónica, la página muestra una advertencia y bloquea la posibilidad de trabajar sobre los datos ya registrados, como se ilustra en la figura 14.

Figura 14. Mensaje de la plataforma sobre un enlace caducado



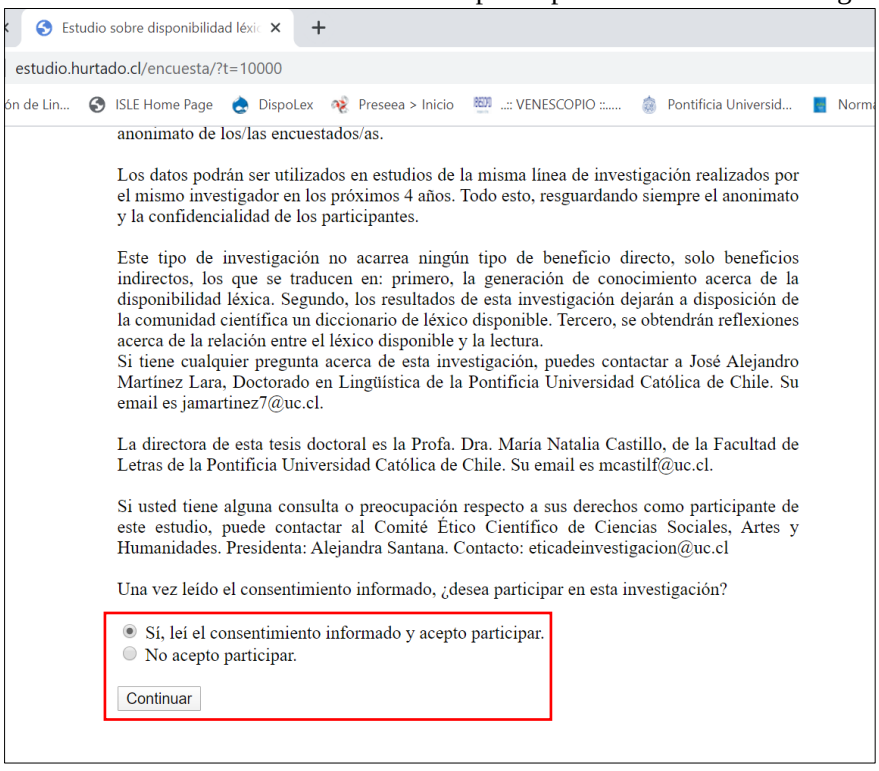
Esta aplicación digital contiene cuatro secciones autónomas, las cuales se relacionan con los instrumentos utilizados en el estudio, en la misma secuencia en la que se administran en el método tradicional. En la primera, se muestra el consentimiento informado, en el que se explica en qué consiste la investigación, los objetivos, los instrumentos y los riesgos y beneficios de la participación, como se observa en la figura 15.

Figura 15. Ventana inicial de la página web: consentimiento informado



Al final del documento, la plataforma muestra dos botones que indican las opciones que pueden elegirse, a saber: desear participar o no en el estudio. Una vez escogida cualquiera de estas alternativas, los encuestados debe pulsar *continuar*, para acceder al primer instrumento de toma de datos. En la figura 16 se aprecia esta descripción.

Figura 16. Botón de selección voluntaria de participación o no en la investigación.



Si el individuo selecciona la opción de participar, la plataforma lo remite a la siguiente sección. Por el contrario, si escoge no realizar las pruebas, la página niega el acceso y redireccionar al sujeto a la salida de la página, como se ilustra en la figura 17.

Figura 17. Mensaje de cierre de la plataforma

En la segunda sección de la plataforma, se encuentra alojada la *encuesta sociológica*. Este instrumento presenta preguntas de selección y respuestas cortas, sobre aspectos sociales y educativos generales, como se indicó en el apartado 2.5. A continuación, en la figura 18, se muestra el encabezado de este aparato analítico. Una vez se ha terminado de contestar las preguntas de este cuestionario, se presiona el botón *continuar* –como se ilustra en la figura 19– que, al cliquear, cierra la segunda sección y lleva al participante a la siguiente, donde se encuentran alojados los test de disponibilidad léxica. El instrumento central de este estudio es la prueba asociativa de disponibilidad léxica, la cual se encuentra en la tercera sección del sitio web. En esta, el participante encuentra primero las instrucciones generales, luego, debe presionar el botón *continuar*, de manera que se habiliten las pestañas de los test sobre léxico disponible, como se aprecia en la figura 20.

Figura 18. Inicio de la segunda sección de la página web: encuesta sociológica

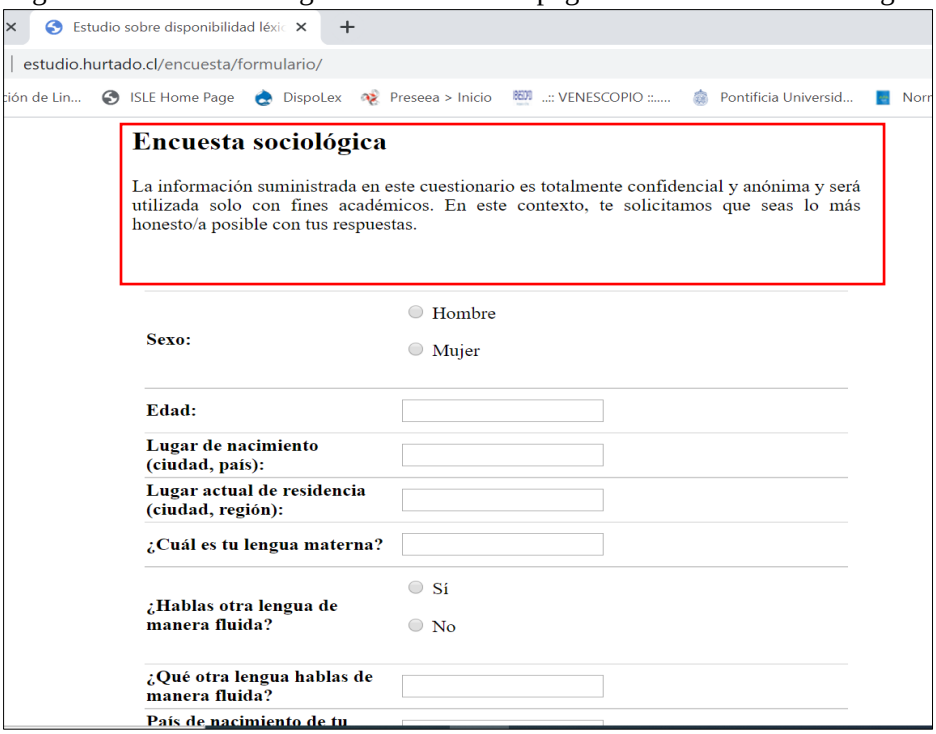


Figura 19. Botón para continuar con la siguiente prueba

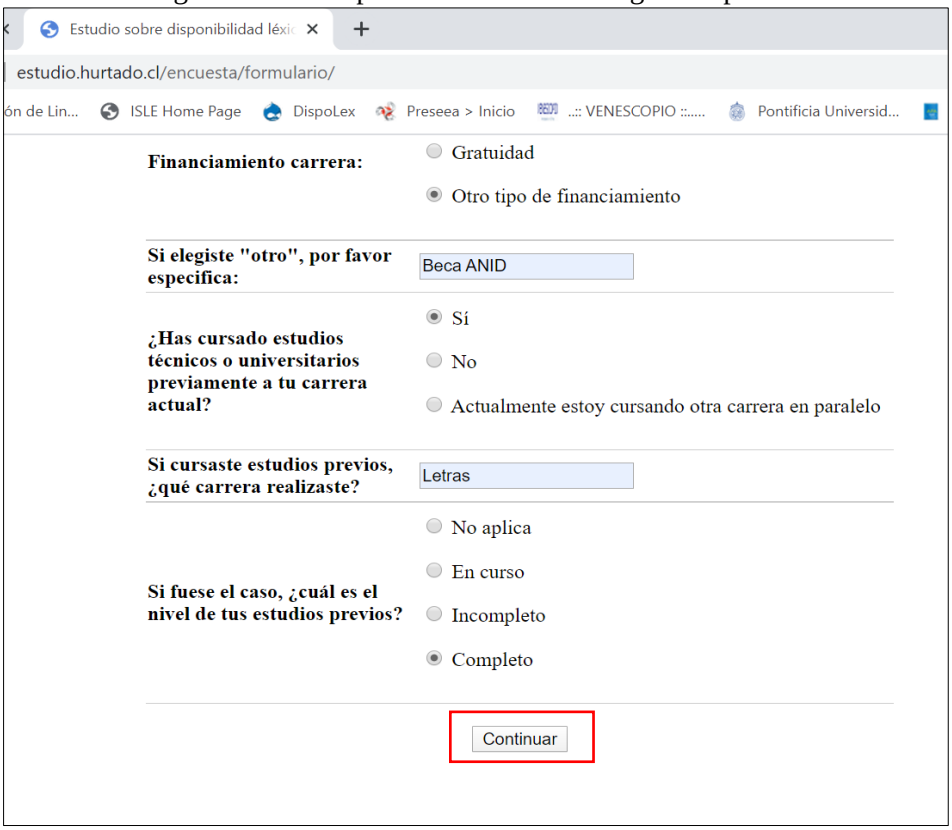
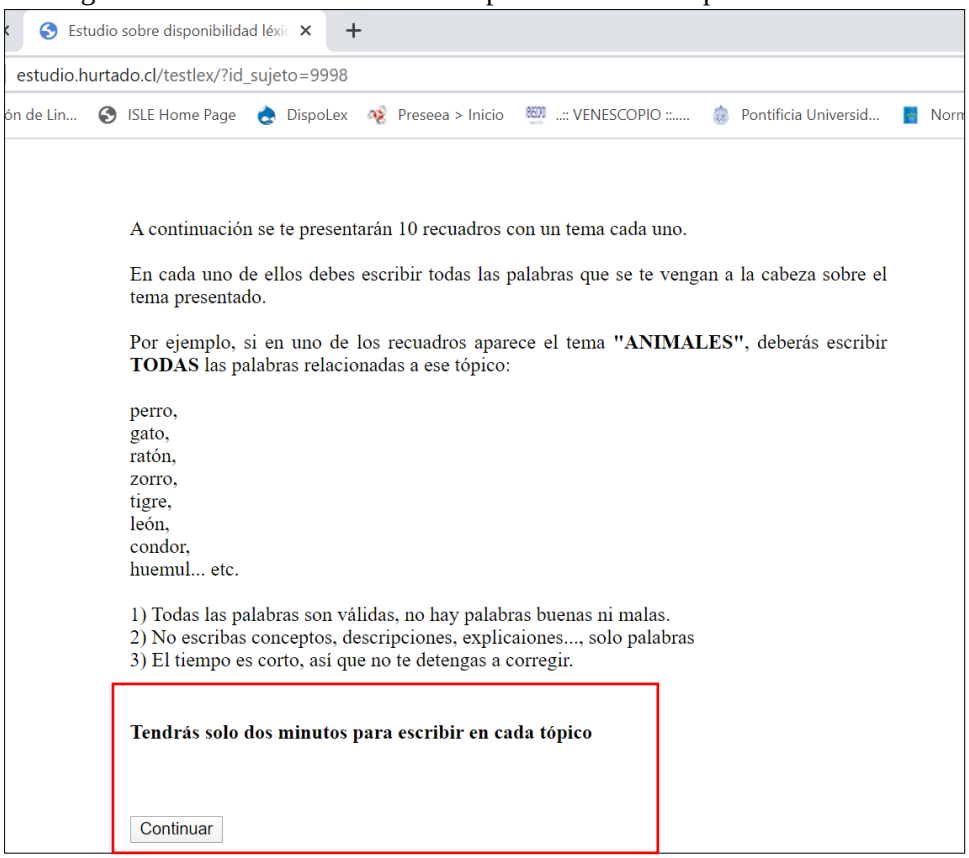
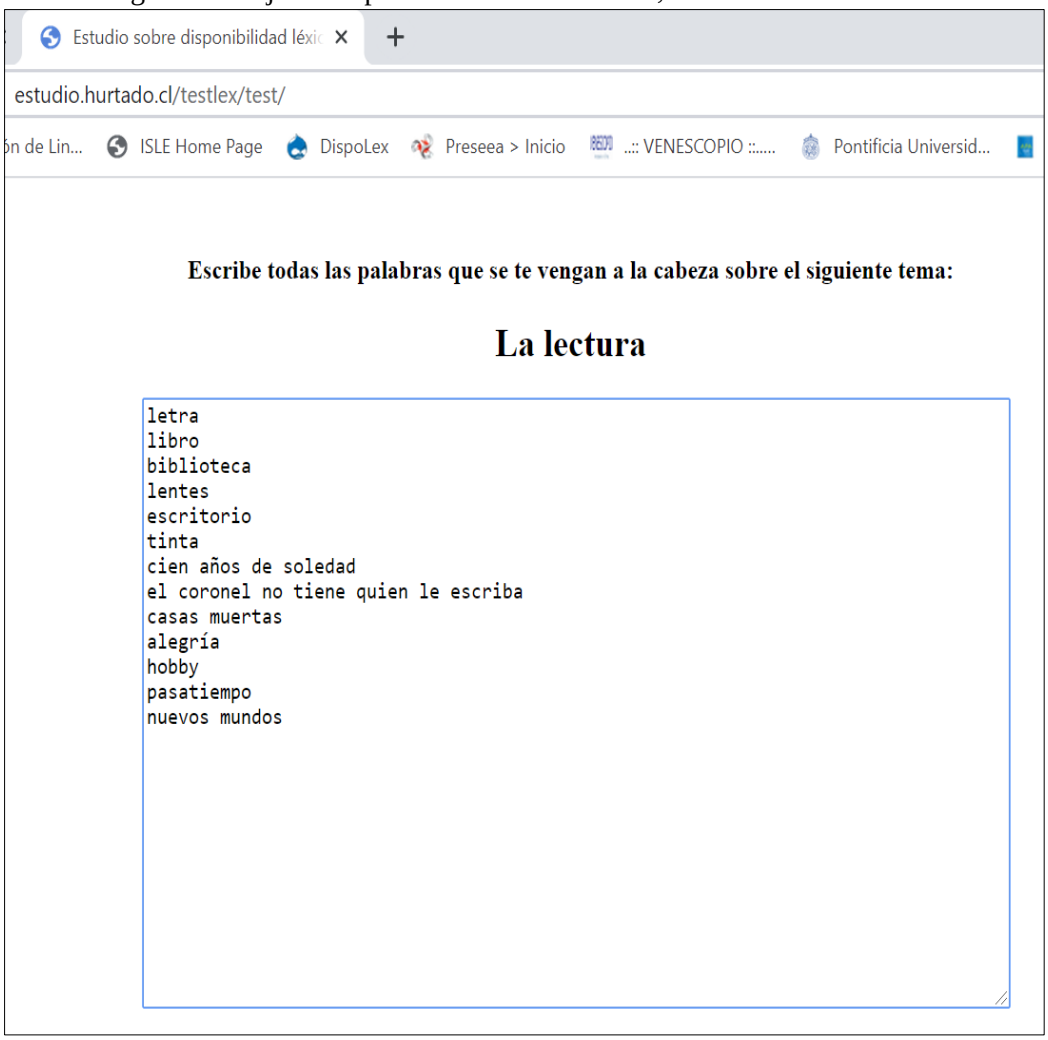


Figura 20. Ventana de instrucciones para los test de disponibilidad léxica



En las ventanas digitales de los centros de interés aparecen: la consigna, el nombre del centro de interés y la caja electrónica de respuestas. Al abrirse el primer cuestionario de DL –en este caso corresponde al eje temático *la lectura*–, empiezan a correr los dos minutos de respuestas. El cronometro no se explicita en la pantalla, por lo que el encuestado no sabe cuánto ha transcurrido desde que inició su escritura. Al terminarse los dos minutos, la caja electrónica se cierra automáticamente. En la figura 21, a continuación, se muestra luna de las hojas electrónicas de respuestas del test de disponibilidad léxica.

Figura 21. Caja de respuestas de los test de DL, actualizador: *la lectura*



Finalmente, después del último actualizador: *comidas y bebidas*, la página web cierra el instrumento de recolección de léxico disponible. Seguidamente, se abre la cuarta sección, donde se halla la batería de prácticas lectoras, que está compuestas por preguntas abiertas, de selección y escala de Likert. El tiempo de respuesta de este cuestionario es libre; por lo que la duración depende del propio encuestado. Cuando el sujeto termina todas las pruebas, clickea el botón *continuar* e, inmediatamente, la página web se cierra absolutamente, se guardan las respuestas en el servidor y el código de acceso caduca.

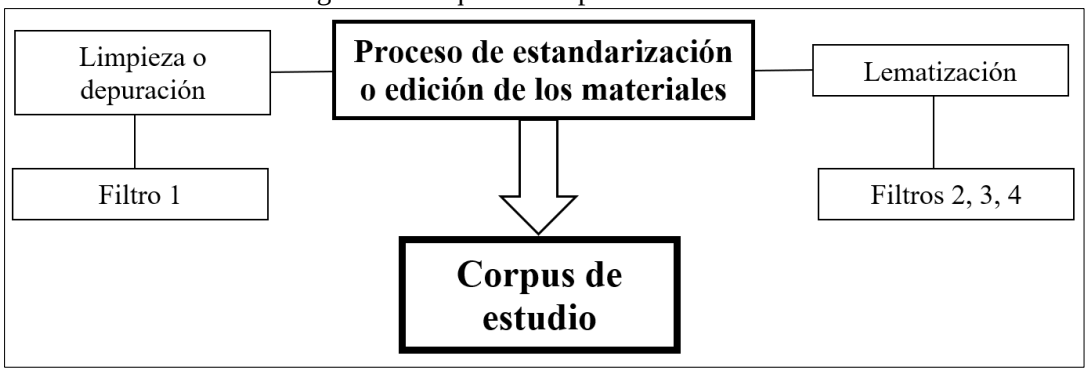
En este subapartado, se han expuestos los instrumentos digitales con los que se recolectaron los materiales léxicos que conforman el corpus de estudios. En el siguiente epígrafe, se detalla la forma cómo se procesó el léxico disponible para los análisis.

2.7. Edición de los datos: depuración y lematización

La intención de este subapartado es exponer la propuesta de edición de los materiales que se ha tomado en cuenta en este estudio, fundamentada en los requisitos manifestados en los trabajos de Samper Padilla (1998), Gómez Devís (2004), Ávila Muñoz (2006), Trigo-Ibáñez (2011) y Santos Díaz (2020). Esta iniciativa no aspira a grandes pretensiones científicas en el campo de la disponibilidad léxica, sino que busca ser práctica, tal como afirma Gómez Devís (2004: 78).

En este texto los términos estandarización y edición de los materiales se emplean como sinónimos. Además, se ha considerado como un protocolo general integrado por dos pasos. El primero, referido a la limpieza o depuración de los listados de palabras; mientras que el segundo consiste en la lematización de las palabras. Esta es un mecanismo metodológico que consiste en "reagrupar las formas heterogéneas de un mismo vocablo, así como separar las formas homógrafas que responden a vocablos diferentes" (Gómez Devís, 2004: 77). En otras palabras, la lematización consiste en homogeneizar todas las expresiones de una misma raíz léxica en una única forma o, lo que es lo mismo, convertir todas las variantes de una palabra en un solo lema, ya que los estudios léxico-estadísticos operan sobre el vocablo o lema que es la unidad léxica básica de lengua (Castillo Fadić y Sologure, 2020). Por esto, la lematización enfrenta problemas complejos (Samper Padilla, 1998: 313). En la figura 26, se esquematiza el sistema de estandarización utilizado en el estudio.

Figura 22. Esquema del proceso de edición



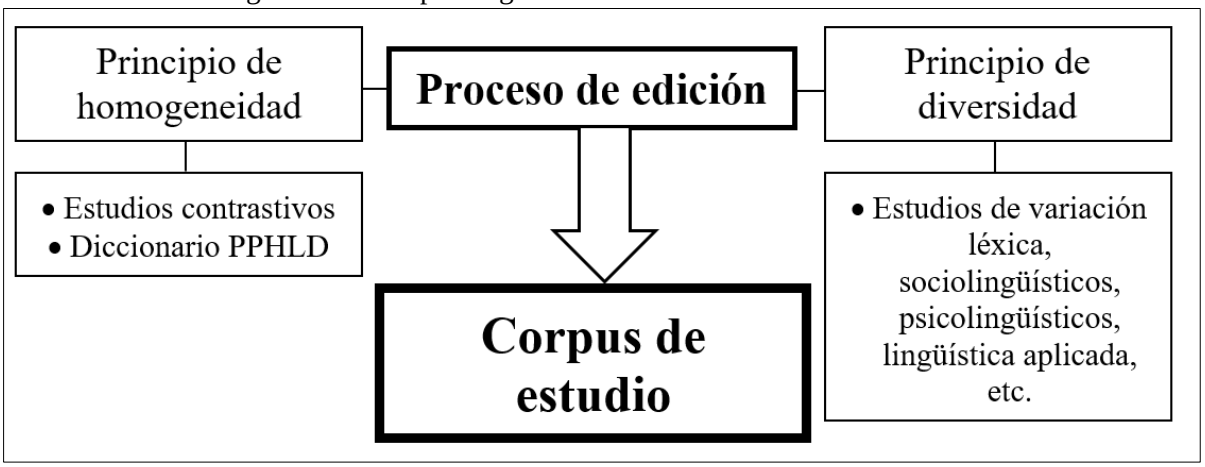
En virtud de las profundas variaciones que existían entre un trabajo y otro en la manera cómo lematizar las piezas léxicas, lo cual imposibilitaba desarrollar comparaciones efectivas, Samper Padilla (1998) expuso un conjunto de criterios y sugerencias con el fin de que los estudios de disponibilidad léxica compartieran un mismo principio de estandarización de las listas de palabras. Estas pautas fueron recogidas e integradas a la metodología del PPHLD. En este sentido, las pesquisas llevadas a cabo en el marco del Proyecto deben seguir la pauta de edición. Sin embargo, el mismo

Proyecto otorga una cierta flexibilidad a los investigadores para que adecuen el tratamiento de los datos en virtud de los objetivos particulares de cada estudio (Ávila Muñoz, 2006: 48).

A pesar de contar con una guía panhispánica, la estandarización de los materiales es uno de los procedimientos metodológicos más controvertidos y complejos dentro del campo de disponibilidad léxica, puesto que debe responder, en primer lugar, al principio de homogeneización de los diferentes corpus; y, en segundo lugar, a las necesidades científicas que se persigue en cada trabajo. Es así como los lexicógrafos se han esmerado en: i) exponer detalladamente las generalidades y particularidades asumidas en la estandarización de las listas de palabras, y ii) proponer sistemas prácticos de edición (Gómez Devís, 2004: 77; Ávila Muñoz, 2006: 53).

Asimismo, el proceso de edición debe seguir dos principios: el de homogeneidad y el de diversidad, los cuales permiten organizar los datos, no solo para los análisis de DL, sino también para los de otros tipos. El principio de homogeneidad consiste en realizar la estandarización de la manera más apegada posible a los criterios generales planteados en el área de la léxico-estadística. En el caso particular del mundo hispánico, se trataría de la propuesta del PPHLD. Esto con el fin de que los materiales léxicos de un grupo puedan ser comparados con los de otros, por ejemplo, entre las variedades del español. Por su parte, el principio de diversidad consiste en la capacidad de lematizar expresiones de manera distinta a lo pautado para el ámbito general, manteniendo una mayor variabilidad lexía en el corpus. La finalidad de esta óptica es: i) conocer más a fondo las unidades léxicas que se distribuyen en un espectro semántico concreto, y ii) realizar estudios desde otras disciplinas y enfoques (Ávila Muñoz, 2006: 48-49). Este planteamiento se ilustra en la figura 23.

Figura 23. Principios reguladores de la edición de los materiales.



Fuente: Elaboración propia a partir de Ávila Muñoz (2006)

En consideración a lo anterior, la delimitación de los protocolos de estandarización suele estar ligada al enfoque teórico asumido. Así pues, desde una perspectiva cognitivista, hay, al menos, dos ópticas: una que se encarga de la conformación de las categorías y otra, de las relaciones asociativas libre (cf. Ávila Muñoz y Sánchez, 2010; Hernández Muñoz, 2006; Hernández Muñoz y Tomé, 2017; Gómez Devís, 2019). La primera toma en consideración que las palabras producidas por los participantes deben pertenecer estrictamente a la categoría léxico-semántica del concepto que activa el centro de interés. Así pues, en la lista de un encuestado X sobre un eje temático Y, si aparecieran palabras no relacionada *stricto sensu* con el área nocional, se excluirían del análisis. Según esto, la siguiente lista de palabras –elaborada por el encuestado 001– debe ser depurada de los vocablos que no refieren directamente a la categoría *Comidas y bebidas*. Por lo que, de las veinte lexías que la componen, solo quedarían trece aptas para el análisis (*papa frita, puré, brócoli, jalea, poroto, lenteja, jugo, Coca-Cola, Sprite, verdura, ensalada, champiñón, arroz*), como se ilustra seguidamente en el ejemplo (03).

(03)

001. papa frita, puré, brócoli, jalea, poroto, lenteja, jugo, Coca-Cola, Sprite, verdura, ensalada, champiñón, arroz

En cambio, la segunda propuesta teórica (asociación libre) permite que se seleccionen todas las unidades léxicas producidas por los encuestados en un área nocional, estén o no vinculadas directamente con la categoría léxico-semántica activada. Desde esta óptica, resulta importante tomar en cuentan las distintas relaciones que tienen las palabras, tanto con otras unidades léxicas de la misma categoría como unidades prototípicamente de otras clases. Gracias a este método, pueden establecerse conexiones de tipo: i) fonética, ii) morfológica, iii) semánticas, iv) pragmáticas, v) experienciales, vi) corpóreas, etc. A manera de ilustración, los lexemas *hambre*, *sed*, *fritura*, *nutrición*, *caloría*, *equilibrio*, *enfermedad*, *cocinar*, *vegetariano*, *vegano*, *contaminación* y *desperdicio* –evocados por el participante 005 en el CI08. *Comidas y bebidas*–, aunque no pertenecen *stricto sensu* a los campos semánticos de la comida y la bebida, se conectan debido a fenómenos de sinonimia, antonimia, experienciales y ecológicos, como en (04).

(04)

005. hambre, sed, fideo, arroz, carne, pollo, pescado, legumbre, fruta, verdura, aceite, fritura, nutrición, caloría, equilibrio, enfermedad, azúcar, cocinar, postre, leche, vegetariano, vegano, contaminación, desperdicio

En esta tesis doctoral, se ha asumido la asociación libre como el abordaje de edición de los materiales. Así pues, se han tomado en cuenta todas lexías reportadas por los encuestados. En los subapartados siguientes, se presentan las pautas específicas que se han aplicado en este procedimiento.

2.7.1. Depuración o limpieza de los datos

La depuración es el paso es previo al de lematización.

2.7.1.1. Criterios comunes con Samper Padilla (1998) y el PPHLD

- *Eliminación de palabras repetidas*. Si una palabra aparecía más de una vez en una misma lista, se suprimía la segunda mención. Este criterio solo se tomó en consideración para el análisis cuantitativo realizado con SPSS y Dispogen. Empero, se mantuvieron (en un documento de respaldo) para el análisis cualitativo mediante Dispografo.
- *Eliminación de marcas gráficas ajenas a la escritura de las palabras*. En algunas ocasiones, los sujetos utilizaron señales como subrayado, asterisco, tachadura, para resaltar algunas lexías.
- *Eliminación de signos de puntuación*, también utilizaban signos como comillas, puntos, signos de exclamación e interrogación, etc.
- *Corrección de las mayúsculas*. Las palabras que estaban escritas, total o parcialmente, en mayúsculas sin que la regla ortográfica actual lo dictaminase, fueron transliteradas en minúsculas.
- *Corrección de los errores ortográficos*, en este trabajo se analiza el léxico de estudiantes universitarios, por lo que se esperaría que este tipo de gazapos no aparecieran. Sin embargo, no están exentos de cometer alguna falta ortográfica, debido a la incidencia de la lengua general como por interferencia de la oralidad, tales como: confusión de grafías, mal uso de tildes y signos diacríticos, pérdida o confusión de letras por razones de fonéticas dialectal.
- *Unificación ortográfica*. Samper Padilla (1998: 314) afirma que no son muchas las palabras que pueden escribirse correctamente de distintas maneras, por ejemplo: *impreso* e *imprimido*. Sin embargo, cuando las hay, debe escogerse una de ellas, ya que, de lo contrario, podrían registrarse dos o más entradas diferentes de una misma lexía, lo que provocaría errores en los análisis léxico-estadístico. En esta investigación, para la escogencia de solo una variante gráfica de una misma expresión, se recurrieron a: Diccionario de la Lengua Española (RAE-ASALE, 2014, actualización del 2022), Diccionario Panhispánico de Dudas (RAE-ASALE, 2005) y Diccionario de Uso del Español de Chile (Academia Chilena de la Lengua, 2010).

a. *Homogeneización gráfica de los neologismos*. La unificación de este grupo de palabras depende del nivel de lexicalización o adaptación lingüística en que se halle la expresión, por lo que puede ser una tarea más complicada. En este trabajo, se siguieron tres pasos. El primero

fue utilizar la forma reportada por el DEL. En segunda instancia, se recurrió a las recomendaciones de otros diccionarios o fuentes lexicográficas confiables, por ejemplo: la base de datos de neologismos del Instituto Cervantes. Respecto a los casos aún no lexicalizados y, por lo tanto, no reportados por los diccionarios, se contrastaron las distintas formas y se seleccionaron las más frecuentes en los motores de búsqueda digital, como Google. Se prefirieron las grafías menos complicadas.

b. *Palabras diferenciadas*, se refiere a los casos en que un concepto se presenta de dos o más formas distintas. Especialmente, en los ejemplos de palabras provienen de una lengua extranjera. Estas pueden tener, al menos, dos formas gráficas distintas, según la manera como el préstamo se haya adaptado, ya sea por calco o por traducción del término, por ejemplo: *fútbol* y *balompié* < *football* (ing.) En estos puntos, se decidió dejar ambas formas como entradas distintas.

c. *Dialectalismos*, se decidió dejar las formas reportadas en los diccionarios, con especial énfasis en el Diccionario de Uso del Español de Chile (2010), ya que los sujetos del estudio son chilenos.

2.7.1.2. Criterios particulares de depuración

- *Eliminación de aclaraciones*. Algunos encuestados hicieron anotaciones sobre ciertas lexías, dichas glosas fueron suprimidas, pero se mantuvieron las unidades léxicas.
- *Uso de mayúsculas*, en este trabajo, al igual que en otros previos (cf. Ávila Muñoz, 2006: 60; Santos-Díaz, 2015: 186), se optó por mantener las mayúsculas en nombres propios, marcas comerciales, abreviaturas, siglas y acrónimos.
- *Mantenimiento de palabras tachadas*. Este criterio priva únicamente para la muestra recogida mediante el método tradicional, puesto que en este los encuestados podían tachar palabras de sus propias listas. Este tratamiento se realizó siempre que: i) pudieran ser reconocidas las lexías a pesar de la enmienda y ii) hayan sido escritas completamente. Se ha decidido esto en virtud de que se parte de la hipótesis de que el nombre del centro de interés activa toda la red asociativa de la categoría y, por ende, cada palabra evocada forma parte de ella.
- *Unidades léxicas idénticas al nombre del centro de interés*, se eliminaron los lexemas que eran iguales a la denominación del área temática, puesto que crea un efecto circular. Sin embargo, se mantuvieron los sinónimos.

2.7.2. Lematización de los datos

2.7.2.1. Criterios comunes con la propuesta de Samper Padilla (1998) y el PPHLD

- *Neutralización de marcas flexivas de género y número*. Los sustantivos y adjetivos que presentaron flexión de género y número se unificaron en sus formas no marcadas. Es decir, en género masculino y número singular. Respecto a los sustantivos heterónimos, se mantuvieron las formas femeninas y masculinas expresadas por morfemas diferenciadores de cada género.
- *Neutralización de la flexión verbal*. Los verbos conjugados se lematizaron en infinitivo. No obstante, se mantuvo la conjugación en los casos de estructuras lexicalizadas.
- *Utilización de paréntesis*. Los paréntesis han servido para indicar la variación (fonética, sintáctica, grafémica) que pueden tener algunas lexías. Además, dan información sobre las estructuras sintagmáticas por medio de las cuales los hablantes pueden referirse a una misma realidad. Los casos en los que se recurrió a estos signos son los siguientes:

a. Acortamientos. Se colocó entre paréntesis la parte truncada. Así, si un sujeto escribía la forma *cole*, referida a 'plantel educativo', esta se lematizó como: *cole(gio)*. Debe acotarse que, en estos casos, no hubo reduplicación de los lemas.

b. Alternancia entre formas plenas y formas reducidas. En otras ocasiones, los hablantes solían reducir a un solo elemento léxico una estructura léxica compleja. Ejemplo: *columna* en vez de *columna vertebral*. En estos casos se rodeó con paréntesis la parte que solía elidirse, como se ilustra a continuación: *columna (vertebral)*.

- *Unificación de las palabras que tienen variación por derivación apreciativa*. Algunas raíces léxicas suelen ser modificadas a través de morfemas apreciativos (diminutivos y aumentativos) sin que esto produzca un cambio en el significado de la palabra. En esto casos, se procedió a registrar la expresión en su forma no marcada. Ejemplo: *perrito* → *perro*. No obstante, se mantuvo el morfema apreciativo, cuando estos implicaban cambio de significado y, por tanto, estaban lexicalizados.

2.7.2.2. Criterios particulares de lematización

- *Inclusión de lexías complejas*. Las unidades plurilexicales –aquellas constituidas por más de una palabra– se registraron y analizaron. Ejemplos: *ser humano* y *carta al editor*.
- *Formas pronominales de los verbos*. Se lematizaron como verbos pronominales los casos que aparecían de esta forma en las listas de palabras. Igualmente, se lematizaron como pronominales los casos que estaban motivados por el contexto al que hace referencia el centro de interés bajo análisis, según las anotaciones de las obras lexicográficas del español. En este sentido, si en las

listas de palabra aparecían las lexías *se gradúa* o *graduarse*, en el centro de interés *La educación*, estas se homogeneizaron bajo la unidad de cita *graduarse*, puesto que, en el contexto educativo, esta forma refiere al 'acto de conferir un grado académico', según el DEL (RAE-ASALE, 2022, versión *online*). En cuanto a los verbos que aparecían tanto en forma pronominal como no pronominal, especialmente los intransitivos que pueden pronominalizarse, se recurrió al uso de paréntesis para indicar esta alternancia. Así pues, *ir* e *irse* se lematizaron como *ir(se)*, sin duplicar los casos.

- *Nombres propios, de obras y de marcas comerciales*. Se registraron los nombres propios de autores, títulos de textos, juegos, videojuegos, marcas comerciales, etc., siempre que estuvieran relacionados con los centros de interés. Se estandarizaron bajo un mismo lema las variantes de los nombres de autores y títulos de obras.

Además de estas pautas generales de lematización, también se hizo necesario definir algunas de carácter especial para cada centro de interés. Esto en virtud de la conformación y naturaleza léxico-semántica y morfológica que pueden tener cada área nocional. En los próximos párrafos, se describen estos criterios.

Lematización del centro de interés La lectura

Según el arqueo bibliográfico, *La lectura* es un centro de interés novedoso. Sin embargo, Ramírez (2019) y Martínez-Lara (2021) han publicado resultados parciales referidos a este actualizador, basados, justamente, en los datos recopilados para esta tesis. La selección de este eje temático se fundamenta en los supuestos de que la población objeto de estudio, por una parte, está en constante práctica lectora, ya que este proceso es indispensable en la atmósfera universitaria. Por otra parte, la lectura es un fenómeno cultura estrechamente ligado con el entorno educativo y profesional de quienes cursan carreras en las facultades de Letras y Educación. En consecuencia, los objetivos de analizar esta área nocional son: i) conocer las palabras disponibles que los informantes relacionan con la lectura como práctica y consumo cultural; y ii) comparar el caudal léxico referido a dicho actualizador que reportan los estudiantes universitarios de dos carreras distintas (Letras Hispánicas y Educación Básica), que comparten intereses sobre la lectura, pero desde sus respectivas perspectivas y enfoques teóricos. En consonancia con los criterios de lematización planteados en esta tesis, se han reservado alguno principio propios de las lexías de este centro de interés, los que se exponen a continuación:

- Se han homogeneizado las variantes de los nombres propios de escritores y obras literarias, manteniendo, como es lógico, el uso de las mayúsculas, según las normas ortográficas del español actual. Ejemplos: Gabriela Mistral, Don Quijote, Alicia, La última fecha, entre otros.
- Con base en el DEL y la Fundéu, se han lematizado en plural los lexemas: lentes, anteojos y medios masivos.
- Se han considerado entradas independientes las siguientes piezas léxicas: comprensión y comprensión lectora, cuento y cuento infantil, divulgación y divulgación científica; literatura, literatura infantil, literatura juvenil; mundo, nuevo mundo, otro mundo, mundo imaginativo y mundo fantástico; obra y obra dramática; tapa, tapa dura, tapa blanda; texto, texto crítico, texto informativo, texto literario, texto no literario; voz, voz alta y voz baja.
- Se han conservado las unidades plurilexicales: ratón de biblioteca y sopa de letras.
- Se ha recurrido a la utilización de paréntesis para indicar que:
  a) la palabra que está dentro de ellos –determinantes– alterna en los listados. Por ejemplo: Fanfic (Tropes), género (literario);
  b) el sustantivo dentro los paréntesis queda sobreentendido, porque refiere al nombre del centro de interés, ejemplo: (lectura) nueva, (lectura) en voz alta, (lectura) atenta; o a una palabra mencionada en la lista, como: libro → (libro) nuevo, (libro) usado, (libro) de bolsillo.
- Se han mantenido las marcas y nombres comerciales de aplicaciones, productos, consumibles y sistemas, algunos de ellos en siglas, tales como: MLA, PDF, iPad, Wattpad, webtoon, plot twist.
- Se ha dejado la palabra monología, que si bien no aparece en las obras lexicográficas consultadas (DLE, DPHD) o en las bases de datos de la Fundéu, sí aparece en textos académicos especializados de literatura y teoría literaria[11].

Lematización del centro de interés El profesor

El eje temático *El profesor* es otro de los centros de interés que, según la bibliografía especializada, no ha sido utilizado en otros estudios de disponibilidad léxica, por lo que uno de los objetivos es aportar datos léxico-estadísticos sobre dicho campo semántico. Respecto al proceso de edición, se tomaron en cuenta las siguientes decisiones para la lematización de algunas formas léxicas.

[11] Para más detalle diríjase a: https://philpapers.org/rec/VERIAL-2

- Las siguientes unidades léxicas se consideraron entradas independientes: ejemplo/ ejemplo a seguir/ ejemplo de vida; figura/ figura de autoridad/ figura (materna/paterna), calidad/calidad humana, clase/ clase virtual, pensamiento/pensamiento crítico, buena onda/mala onda; voz/ voz alta/ voz de mando.
- Con base en el DEL, se lematizaron en plural los vocablos: gafas y padres.
- Se mantuvo la flexión de género femenino en el vocablo: palabrería, según aparece en el DEL.
- Se ha recurrido a la utilización de paréntesis para indicar que la palabra dentro de ellos alterna en el corpus: bata (blanca), figura (materna/paterna).
- Se mantuvieron piezas léxicas complejas, tales como: levantar la mano, sacarse la chucha, manejo de contenido, perfil docente, ser humano.

Lematización del centro de interés La educación

Este actualizador resulta novedoso en los estudios de disponibilidad léxica. Fue utilizado por primera vez por Herranz (2018, 2020), bajo el término: Educación. Luego, se han publicado algunos resultados sobre esta misma área nocional en el trabajo de Martínez-Lara (2021). En esta investigación el empleo de este centro de interés contribuye con el conocimiento léxico-estadístico de este eje temático. Las particularidades de la edición de los casos son las siguientes:

- Se han registrado como lemas distintos las unidades léxicas: clase/ clase social, lucha/ lucha de clase, matemático/ matemáticas, política/ políticas públicas, pensamiento/ pensamiento crítico, tía/ tía del aseo, transmisión/ transmisión de conocimiento, viaje/ viaje de intercambio, práctica/ práctico, secundaria/secundario, sala (de clase) / sala musical, útil/ útiles, sistema/ sistema deficiente/ sistema online.
- Se conservaron las lexías complejas: ciencias sociales, ciencias naturales, sistema online, clima educativo, centro de alumnos, Centro de Formación Técnica, conexión neuronal, cuarto medio, ensayo y error, entorno social, lugar seguro, mal pagado, mamá gallina, pedagogía Waldorf, pérdida de tiempo, reinserción social, sentido de vida, jornada escolar.
- Se mantuvo la flexión de género en los casos requeridos, según la RAE (2014) o determinado por el cotexto: práctica, secundaria, (educación) media, matemáticas.
- Se mantuvo la flexión de número los vocablos que lo ameritaban, en consonancia con el DLE: humanidades, padres, útiles, lentes, matemáticas, modales.
- Se utilizaron los paréntesis para indicar la variación de las lexías: sala (de clase), (educación) formal, (educación) informal, (educación) no formal, jardín (infantil).

- Se mantuvieron las siglas y los acrónimos: CAE, Junaeb, PPT, PSU, PTU, PAES.
- Se conservaron los nombres de programas digitales: Canvas, Power Point, Word, Zoom.

Lematización del centro de interés Juegos y distracciones

Este centro de interés forma parte del grupo de los tradicionales. Su objetivo es conocer el vocabulario asociado a las actividades recreativas de los encuestados. Además, se busca poder comparar los resultados de este estudio con otros previos. En el proceso de edición de los materiales se tomaron las siguientes consideraciones particulares para esta área nocional:

- Los siguientes vocablos han sido registrados como entradas diferentes: relación/ relación sexual, saltar/ saltar la cuerda, silla/ silla musical, cama/ cama (elástica/ saltarina).
- Se mantuvo la flexión de número en las palabras: manualidades, rompecabezas, como se indica en el DEL.
- Se mantuvo la flexión de género en el vocablo muñeca.
- Los vocablos con variación léxica se identificaron a través de paréntesis: (página) web, (juego) en línea, cama (elástica/ saltarina), bici(cleta), compu(tador/a), super(mercado), tele(visión), escuchar (música), Play(Station), tiempo libre, TV (cable).
- Las lexías complejas se dejaron en los listados: aire libre, banda ancha, caballito de agua, cara de caca, carrera de saco, centro comercial, cortar revista, juego de mesa, mesa de billar, montaña rusa, mono animado, no estudiar, parque de diversiones, pérdida de tiempo, realidad virtual, simulador de vuelo, sopa de letras, subir cerro, tenis de mesa.
- Se han mantenido los nombres propios de los juegos, tanto los reportados en español como e inglés: Calabozos y Dragones, Free Fire, League of Lengends, The Lengend o Zelda, Mario Bross, PUBG, Uno, Lego, Alice: Madness Returns, American McGee's Alice, Adivina quién?
- Se han registrado los nombres de aplicaciones, páginas de internet y redes sociales: Facebook, Instagram, Netflix, Google, YouTube, Amazon Prime.
- Se han registrado como entradas distintas las siglas de los nombres de juegos, programas, aplicaciones e instituciones deportivas y recreativas: LOL, COD, FIFA, PUBG.
- Se han mantenido las unidades léxicas que expresan alguna experiencia sobre este centro de interés: buen rato, chao estrés, fin de semana, tiempo libre, ruptura de la rutina, mal humor, mala leche, matar el tiempo.
- Se tildó el vocablo vóleibol, según la forma adoptada en Chile y registrada en el DEL.

Lematización del centro de interés La escuela: muebles y materiales

Uno de los dieciséis centros de interés propuesto por Gougenheim *et al.* (1964) y aplicado en los trabajos del PPHLD es *La escuela: muebles y materiales*. Con este se ha querido no solo recoger las unidades léxicas que hacen referencia a los elementos físico de los establecimientos educativos, sino también se ha querido conocer que otro tipo de evocaciones vienen a la mente de encuestados, quienes están compenetrados en el sistema educativo. Los criterios particulares de edición fueron:

- Se registraron como lexías distintas: aire/ aire acondicionado, caja/ caja de arte, clase/ clase social, clase alta/ clase baja, diario/ diario mural, dispensador de alcohol/ dispensador de jabón, espacio/ espacio seguro, goma (de borrar) / goma Eva, mesa/ mesa de profesor, papel/ papel crepé/ papel de volantín/ papel higiénico/ papel kraft/ papel lustre, lápiz/ lápiz de color / lápiz gel/ lápiz mina/ lápiz pasta, sala (de clase) / sala de música/ sala de práctica, útiles/ útiles de aseo.
- Se registraron dentro de paréntesis y separadas por barra oblicua las palabras educación y formación como locativos de centro, porque las mencionó el mismo informante: centro (educación/formación).
- Se registraron como entradas distintas las siguientes palabras porque la flexión de género las diferencia: banco/ banca, gimnasio/ gimnasia, pizarra/ pizarrón.
- Se mantuvieron las lexías complejas: alcohol gel, aro de básquetbol, carné de biblioteca, corta cartón, crema de mano, elemento de práctica, juego de mesa, mesa del profesor, palo de helado, lectura veloz, libro de notas, vuelta a clase.
- Se mantuvo la flexión de número en los vocablos que lo ameritaban, en consonancia con el DLE: útiles, burdeos, tiralíneas.
- Se registraron los nombres de programas computacionales, de materiales y marcas comerciales: Zoom, Liquid Paper, Bip!, Confort, placa de Petri, Power Point.
- Se registraron también las siglas y los acrónimos: CRA, Bic, Ceneval, PPT.
- Se mantuvieron los morfemas apreciativos en las palabras cuadernillo, manilla, ya que son usos lexicalizados.

Lematización del centro de interés Habilidades y cualidades docentes

Entre los centros de interés nuevos, y del que no hemos publicados resultados preliminares, se encuentra *Habilidades y cualidades docentes*. Este, además de representar un aporte a los estudios de disponibilidad léxica al ampliar los conocimientos acerca del caudal léxico del grupo encuestado, tiene el objetivo específico de identificar las palabras que los estudiantes de pedagogía utilizan para

expresar las nociones que tienen sobre la profesión docente. En el proceso de edición de los materiales, se tomaron las siguientes decisiones:

- Se mantuvieron las unidades plurilexicales: actitud crítica, agilidad mental, aguantar pipí, amante del progreso, buen corazón, buena onda, buena persona, buscar herramienta, ciencias naturales, emplear recursos, espacio seguro, estabilidad mental, estar de pie, traspasar conocimiento, manejo de grupo, manejo de conocimiento.
- Se consideraron entradas diferentes: competencia/ competencia (meta)cognitiva/ competencia cultural, comprensión/ comprensión lectora, expresión/ expresión corporal, ganas/ ganas de ayudar/ ganas de enseñar, hablar/ hablar claro/ hablar correctamente/ hablar fuerte, lenguaje/ lenguaje (verbal/ no verbal), transmitir/ transmitir conocimiento, salud/ salud (física/ mental).
- Se registraron dentro de paréntesis y separadas por barra oblicua las palabras verbal y no verbal, física y mental, por ser locativos de lenguaje y salud, respectivamente. Estas fueron mencionadas por el mismo informante: lenguaje (verbal/ no verbal), salud/ salud (física/ mental).
- Los vocablos con variación léxica se identificaron a través de paréntesis: expresión (corporal), herramientas (pedagógicas).
- Se mantuvo la flexión de número en el vocablo *manualidades*, *matemáticas*.

Lematización del centro de interés Partes del cuerpo

El séptimo centro de interés analizado en este trabajo es *Partes del cuerpo*, que forma parte de los ejes temáticos tradicionales en los estudios de disponibilidad. Contar con este actualizado garantiza el contraste con otros grupos. Las particularidades de la lematización en este campo fueron:

- Se mantuvieron los dialectalismos (chilenismos) que refieren a alguna parte del cuerpo: pera, pupo, pechuga, pico, pingüino emperador, ñata.
- Se distinguieron los lemas: órgano/ órgano sexual, dedo/ dedo chico, glándula/ glándula adrenal/ glándula mamaria/ glándula sudorípara; glóbulo/ glóbulo blanco/ glóbulo rojo; intestino/ intestino grueso/ intestino delgado.
- Se mantuvieron los eufemismos: partes íntimas, mirada de amor.
- Se registraron entre paréntesis las partes variables de las unidades léxicas: columna (vertebral), (zona) lumbar, color (de piel), cuerda (vocal), fosa (nasal), médula (espinal).
- Se conservó la flexión de número en los siguientes casos: bronquios, extremidades, bíceps, genitales, tríceps,
- Se registró la sigla ADN.

- Distinción entre español e inglés: bubi y booby

Lematización del centro de interés Comidas y bebidas

Por último, se lematizaron las unidades léxicas correspondientes al centro de interés *Comidas y bebidas*, que forma parte también de los actualizadores tradicionales. En este se encontraron distintos términos que necesitaron un tratamiento particular de edición. Los criterios considerados fueron:

- Se mantuvieron los nombres de platos y preparaciones: arrollado primavera, carne al jugo, carne asada, carne mechada, filete a lo pobre.
- Se registraron los dialectalismos (chilenismos): anticucho, poroto, queque, sopaipilla, completo.
- Se mantuvieron los nombres de las cadenas de restaurantes nacionales: Juan Maestro; Pedro, Juan y Diego; Ramón; Dulcepan.
- Se encerraron entre paréntesis y separaron con barra oblicuas las dos formas en las que se denomina una unidad léxica, manteniendo fuera del paréntesis la parte invariable: jugo (de sobre/en polvo).
- Se mantuvieron entre paréntesis las formas alternantes: (comida) callejera, (comida) chatarra, compa(ñero), mayo(nesa), (bebida) gaseosa.
- Se distinguieron como entradas distintas las formas: agua/ agua mineral/ agua saborizada/ agua con gas, hamburguesa/ hamburguesa de soya, castaña/ castaña de cajú, diente/ diente de dragón, harina/ harina de maíz/ harina de trigo, carne/ carne de soya, huevo/ huevo duro/ huevo frito/ huevo revuelto.
- Se mantuvo en minúscula el vocablo bon o bon, ya que así está registrado el nombre comercial.
- Se dejó en mayúsculas el nombre de la marca KAPPO, pues así está registrada.
- Respecto al uso de diminutivos, se ha dejado la forma guatita, que refiere al plato a base de panza de cerdo.

2.8. Codificación de los casos

Una vez que se han descritos los instrumentos y los métodos de recolección de los materiales, se reseña ahora la manera cómo fueron preparados y codificados para los análisis cuantitativos y cualitativos. Se entiende por codificación el procedimiento mediante el cual se rotulan las características –en este trabajo, sociológicas y socioeducativas– de cada caso en una secuencia numérica, ordenada y biunívoca, de tantos caracteres como variables de estudios haya (Martínez-Lara,

2012). Este proceso es necesario para los análisis mediante los programas computacionales: Dispogen, Dispografo e IBM SPSS®.

Respecto a Dispogen, esta es una aplicación creada por Max Echeverría y su equipo, en 2005, con la finalidad específica de analizar el léxico disponible de un grupo. Para su elaboración se recurrió al programa MatLab. Así pues, Dispogen calcula los índices generales y específicos de DL, a saber: el número de palabras (NP) y vocablos (NV), el promedio de palabras (PP o $\overline{X}$), el índice de cohesión (IC) y, el más importante, el índice de DL (IDL). Este último está basado en la fórmula de López Chávez y Strassburger Fría de 1987, corregida en 1991, ver subapartado 1.7. Sobre el índice de cohesión, a grandes rasgos, se refiere al cómputo que índica el nivel o grado de compactibilidad de las piezas léxicas dentro de un centro de interés particular (Echeverría *et al.*, 1987).

Por su parte, Dispografo es un programa computacional diseñado también por Max Echeverría y su equipo de la UdC, en el marco de las investigaciones de DL. El objetivo de esta aplicación es mostrar las asociaciones de las palabras disponibles de un CI a través de grafos. Estos exponen las imbricadas redes construidas por nodos, con sus respectivas aristas, que se determinan por medio de un algoritmo que lee las secuencias de las lexías. Una de las finalidades de este software es colaborar con los análisis psicolingüísticos (Echeverría *et al.*, 2008).

Por último, SPSS es una aplicación del paquete estadístico de IBM® que ofrece un conjunto de pruebas estadísticas, como t de Student, Anova, Chi cuadrado, entre otras, las cuales fueron requeridas para los análisis complementarios de este trabajo.

En la codificación fue necesario contar, además, con el programa Excel®, puesto que brinda ayuda en la inserción, organización y sistematización del universo de los casos. A diferencia de los textos planos, Excel® cuenta con sistema de filtro y fórmulas que permiten seleccionar, codificar y recodificar de manera fácil y automática los materiales y, además, ayuda a detectar los errores de la codificación (Mangado y Areta, 2008: 481). En este sentido, en un documento Excel® se vaciaron todos los casos de los encuestados de manera organizada. En la primera columna de una hoja Excel, se colocaron los códigos identificadores (ID) de los participantes y, en las subsiguientes, toda la información suministrada por ellos (una pregunta o ítem por columna). Así, cada fila corresponde a un único individuo. Este sistema permite, extraer toda la información pertinente para el análisis computacional y el contraste de las hipótesis o preguntas de investigación (Herrera *et al.*, 2011).

2.8.1. Datos sociológicos y de prácticas lectoras

Si bien los datos sociológicos y los de prácticas lectoras se tomaron a través de instrumentos distintos, las respuestas de ambas encuestas se transliteraron en códigos numéricos, en dos pasos. En el primero, se registraron todos los valores originales en una hoja. Por ejemplo, para la pregunta referida a *sexo*, se guardaba la respuesta (H para hombre y M para mujer). En el segundo paso, en una hoja Excel nueva, se pasó esta información a caracteres numéricos. Así pues, para el caso de la respuesta H (hombre), se adjudicó el número 1, mientras que para M (mujer), el 2. Vale acotar que esto se hizo únicamente con los ítems categóricos o de preguntas cerradas.

Este proceso de codificación permite realizar distintos análisis estadísticos, especialmente lo de corte no paramétrico (Herrera *et al.*, 2011) a través de SPSS. Además, como se ha indicado en los párrafos previos, Excel® también es ideal para la organización y selección de las variables que se exportan a Dispogen y Dispografo; en estos casos, después de haber convertido la información pertinente en formato .txt o texto plano.

Una vez realizada la codificación, se procedió a determinar la naturaleza de cada variable de estudio, lo que depende del nivel de exactitud informativa de los datos recogidos. Así pues, la bibliografía refiere a distintos niveles de medida de exactitud, tales como: nominal, ordinal, intervalo y razón (Mafokozi, 2009). Sin embargo, en esta sección, se explican los dos primeros que son los que corresponden a las variables de este trabajo.

- Nivel nominal: se refiere a los valores que ayudan a distinguir a los objetos, en función de sus características, en el que uno tiene un atributo que no posee el otro (Mafokozi, 2009; Herrera *et al.* 2011). Por ejemplo: *carrera*. En este caso, las respuestas podían ser: *Educación Básica* o *Letras Hispánica*, por lo que es una variable dicotómica. Entonces, a la primera se le asignó el dígito 1, mientras que, a la segunda, el 2.
- Nivel ordinal: en este caso, los valores asignados a cada nivel, además, de distinguirse uno del otro, también reflejan el orden o jerarquía que tienen con respecto al factor (Mafokozi, 2009; Herrera *et al.* 2011). Por ejemplo: *Frecuencia de lectura optativa* tienen cuatro niveles o variantes: *Ninguna hora*; *De 1 a 5 horas*; *De 6 a 10 horas*; *Más de 10 horas*, por lo que es una variable politómica. Entonces, a cada una se les asignó un número del 1 al 4, en el que 1 representa el nivel más bajo y el 4, el más alto. Así pues, el código 3 no solo funciona para distinguir el nivel 3 del 4, sino que además sirve para indicar que es menor a este último.

Una vez descrito el tratamiento de los datos sociales y de prácticas lectoras, se explica el procesamiento de la información léxica.

2.8.2. Datos léxicos

Para las ocho listas de palabras elaboradas por cada encuestado, el tratamiento difiere completamente de los sociológicos y de lectura, ya que, en este caso, a las palabras no se les asignó un valor número que representara un código. Por el contrario, ellas se dejaron tal cual, según los criterios de edición establecidos y descritos en el subapartado 2.7. A fuer de, Santos Díaz (2020: 108) afirma que "el objetivo no es codificar una variable para interpretarla, sino preservar o alterar las palabras que han escrito los informantes atendiendo a criterios de edición". Entonces, a pesar de que las listas de palabras corresponden a la variable dependiente del estudio, estas no son rotuladas numéricamente, porque no se interpretaron estadísticamente cada una como variantes o elementos alternantes de cada centro de interés, lo que cuantitativamente sería inabordable.

2.8.3. Plan de codificación

En consonancia con los puntos tratados en los apartados anteriores, en este epígrafe se explica el plan de codificación de las variables del estudio, que toma en cuenta los objetivos de la investigación, las características de los programas computacionales utilizados para realizar los cálculos y la naturaleza de las variables independientes. En este sentido, resulta elemental, primeramente, acotar que se entenderá por variable y variante, en función del diseño de la investigación.

En palabras de Herrera *et al.* (2011: 20): "Por variable entendemos cada rasgo o característica que se registra en una muestra de individuos, que siendo común a un cierto grupo de individuos u objetos tienen distintos grados de magnitud". Entonces, una variable es una categoría que agrupa elementos que comparten algunos atributos comunes. Por su parte, las variantes de una variable corresponden a las características o niveles distintivos que componen la variable. Así, una muestra constituida por individuos de una misma ciudad puede presentar –entre otras– dos variables distintas como: *estratos socioeconómicos* y *grupo generacional*. En el caso de la primera, los sujetos podrían ser de: *clase alta*, *media* o *baja*; mientras que, por la segunda, podrían estar en los rangos: *de 20 a 30 años*, *de 31 a 40 años*, *de 41 a 50 años* o *más de 50 años*.

Para este estudio, se han considerado seis variables independientes –las que permiten observar si existen cambios en la variable dependiente (López Morales, 1994b; Herrera *et al.*, 2011)–, a saber: 1) *sexo*, 2) *carrera*, 3) *año o nivel de curso*, 4) *formato de las pruebas*, 5) *cantidad de libros leídos* y 6) *frecuencia de lectura optativa*. En la tabla 9 se ilustra el plan de codificación de las variables

independiente con los códigos específico de cada variante –los que se usaron para los cálculos en los programas computacionales–, además, de la información sobre el nivel y tipo.

Tabla 9. Plan de codificación de las variables independientes

Variables independientes	Variantes	Código	Nivel	Tipo
Sexo	Hombre Mujer	1 2	Nominal	Dicotómica
Carrera	Educación Básica Letras Hispánicas	1 2	Nominal	Dicotómica
Año de curso	1.er. año 4.o. año	1 2	Ordinal	Dicotómica
Formato de la prueba	Papel Digital	1 2	Ordinal	Dicotómica
Cantidad de libros leídos	Ningún libro De 1 a 5 De 6 a 10 Más de 10	1 2 3 4	Ordinal	Politómica
Frecuencia de lectura optativa	Ninguna hora De 1 a 5 horas De 6 a 10 horas Más de 10 horas	1 2 3 4	Ordinal	Politómica

2.8.4. Procesamiento de los datos

Para este paso se atendió a los siguientes tres parámetros (cf. Ávila y Lasarte, 2010).

Parámetro 1. Después de que las listas de palabras fueron, primero, vaciadas en la base de datos en Excel®; segundo, editadas (depuradas y lematizadas), según los principios preestablecidos; por último, convertida la información sociológica y de prácticas lectoras en números, de acuerdo con el plan de codificación, se procedió a rotular cada lista de palabras. Luego, se convirtieron los documentos en texto plano (.txt), que es el formato requerido por los programas Dispogen y Dispografo; en la Figura 24 se muestra la pantalla de edición y registro de las categorías a analizar en Dispogen. Estas aplicaciones permiten introducir únicamente cinco variables, que deben tener los siguientes rasgos (Figura 25):

- Las variables 1, 2 y 3, deben ser de dos variantes, como *sexo*.
- La variable 4 debe tener 3 niveles, como *tipo de secundaria*.
- La variante 5 debe tener 4 variantes, con *cantidad de libros leídos*.

Figura 24. Ventana para la edición de las categorías de análisis en Dispogen

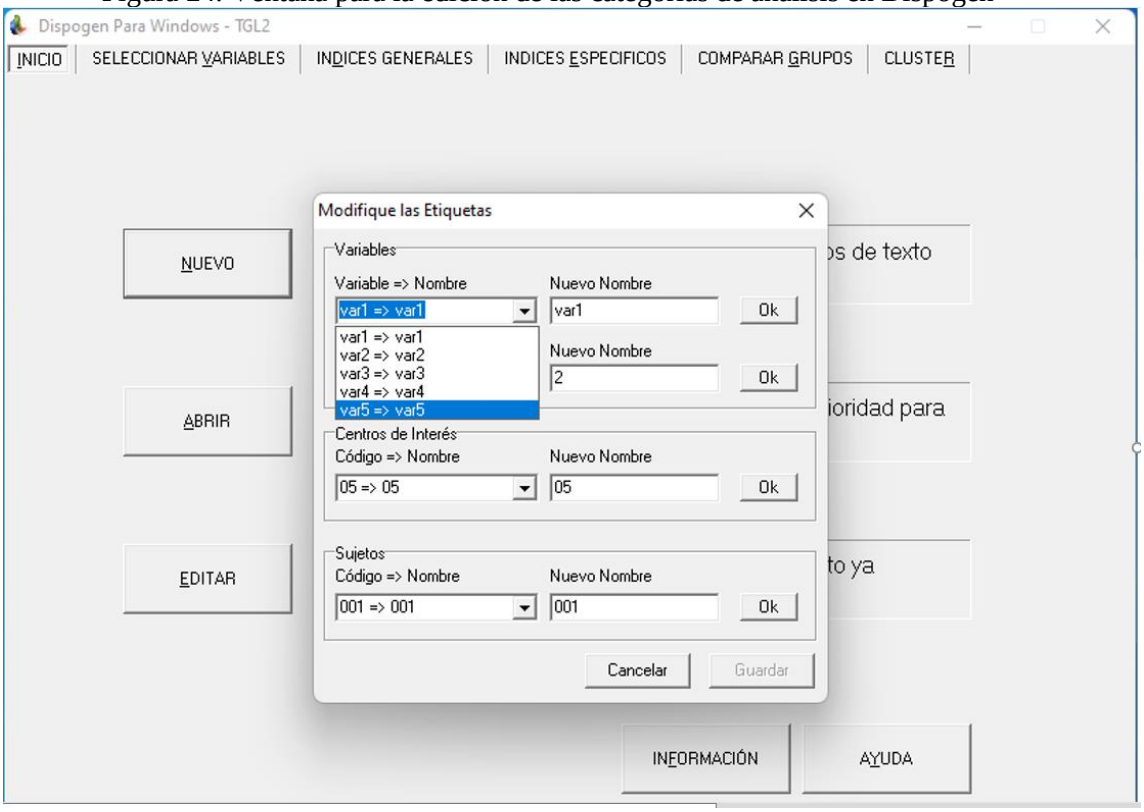
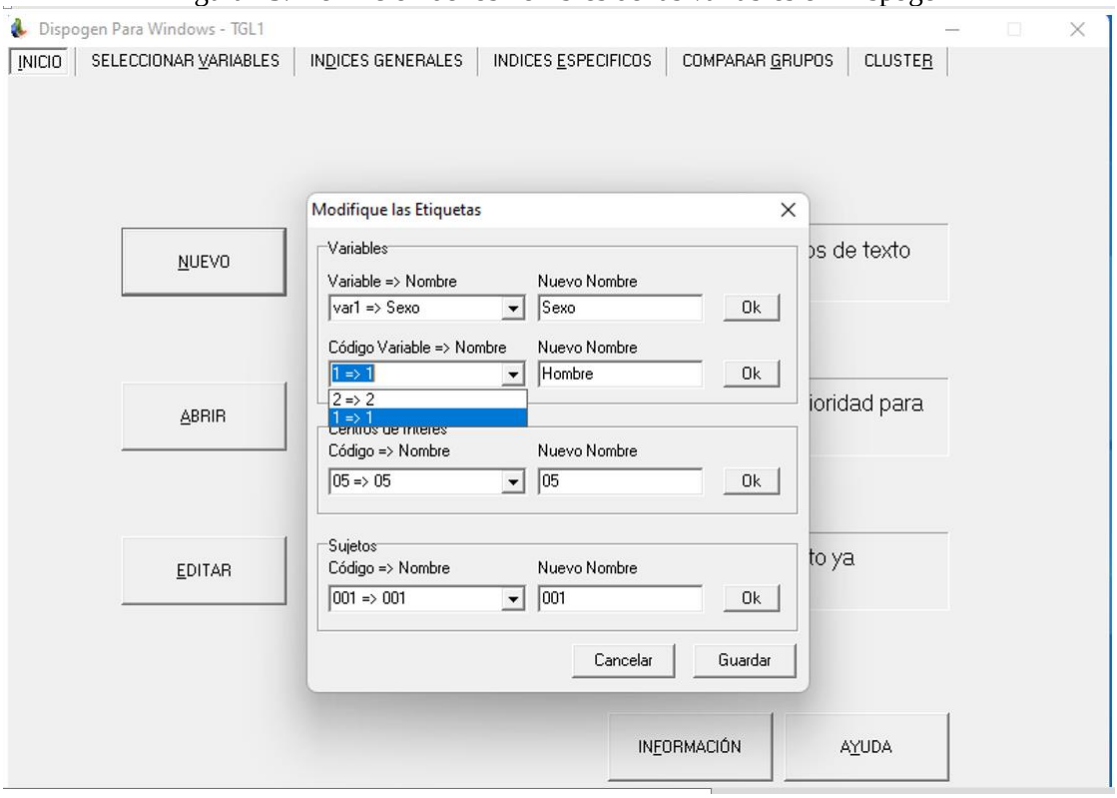


Figura 25. Definición de los nombres de las variables en Dispogen



Asimismo, deben rotularse también los participantes y los centros de interés. En el caso de los sujetos, Dispogen y Dispografo admiten hasta 999 encuestados, por lo que la codificación de estos debe contar con tres dígitos. Por su parte, los centros de interés deben ser registrados con dos dígitos. En esta investigación los centros de interés son ocho, por lo que la numeración va del 01 al 08, como puede apreciarse en la tabla 10.

Tabla 10. Código de los centros de interés analizados.

Centro de interés	**Código**
La lectura	01
El profesor	02
La educación	03
Juegos y distracciones	04
La escuela: muebles y materiales	05
Habilidades y cualidades docentes	06
Partes del cuerpo	07
Comidas y bebidas	08

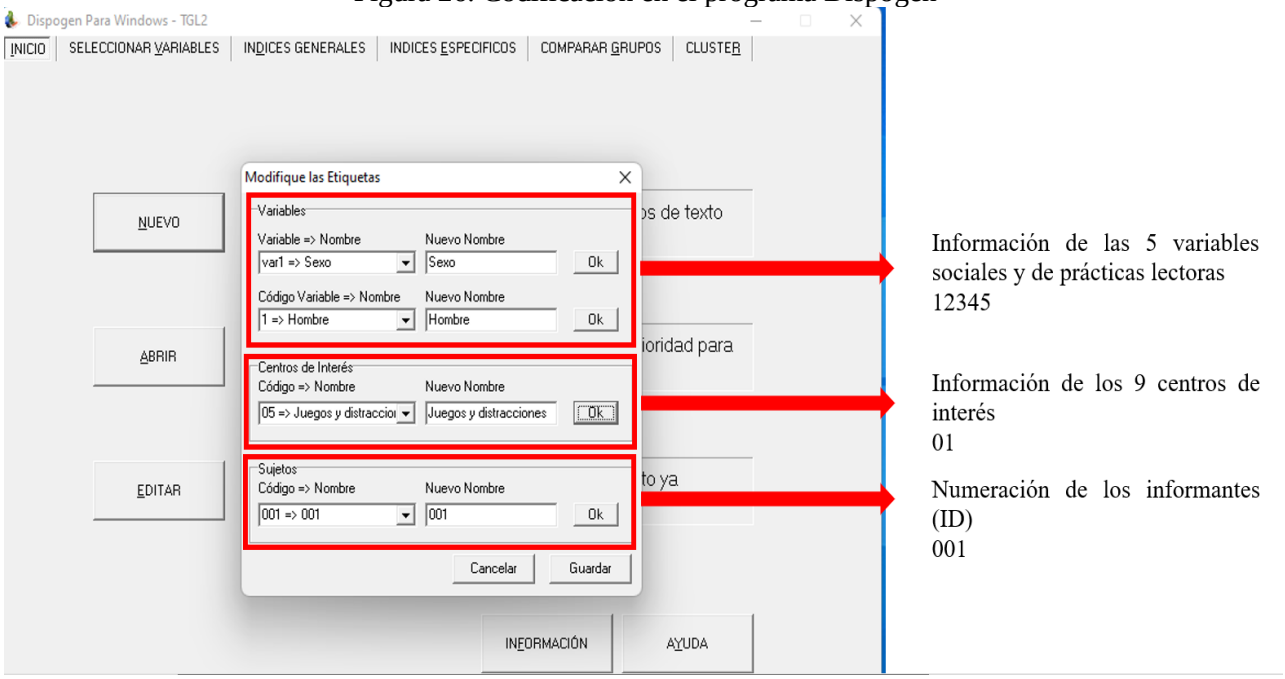
En este marco, la codificación de los datos está organizada en tres partes. En la primera, se presentan los dígitos de las cinco variables extralingüísticas. En la segunda se expone el ID del encuestado, conformado por tres números. Por último, la tercera exhibe los dos números que identifica al centro de interés de la respectiva lista. Cada una de estas tres partes del código está separadas por un espacio. En (5) se ejemplifica la estructura de la codificación de una lista de palabras.

(5)

12345 123 12 palabra1, palabra2, palabra3, palabra4…

En la figura 26 se representan las tres partes de la codificación en el programa Dispogen.

Figura 26. Codificación en el programa Dispogen



Parámetro 2. Los programas Dispogen y Dispografo admiten únicamente documentos en texto plano o .txt, además, las palabras deben estar separadas por coma. Este formato, denominado Salamanca, permite que las aplicaciones lean y procesen las piezas léxicas entre comas, aunque la conformen dos o tres elementos (ejemplo: *buena onda*), como un único vocablo. En este sentido, si en una secuencia de palabras codificadas como: *12345 123 12 puerta*, *ventana*, *techo*, *ojo de buey* … Esta última lexía (*ojo de buey*) se entenderá como un vocablo. En la Figura 27 se ilustra este punto.

Parámetro 3. La lematización de las palabras permite el uso de mayúsculas, algunos signos de puntuación (paréntesis, guiones cortos, barras oblicuas); la grafía ñ; palabras compuestas, ya que el programa puede leerlos.

Los datos de este estudio fueron procesados y codificados en atención a estos parámetros. En este orden de ideas, si ingresaron a Dispogen los datos de las variables: *sexo*, *carrera*, *nivel de curso*, *formato de la prueba* y *cantidad de libros leídos*. Así pues, el código 11234 023 03 significa que las palabras de la lista fueron escritas por: i) un hombre, ii) de Educación Básica, iii) cursa cuarto año, iv) el formato de la prueba es digital, y v) ha leído más de diez libros en el año. Además, ocupa el puesto 23 de los 264 de los informantes y las piezas léxicas corresponden al tercer centro de interés: *la educación*.

Figura 27. Ilustración de una secuencia de palabras en formato Salamanca y documento .txt

```
21111 001 01 libro, concentración, historia, lenguaje, atención, destacador, book
21113 002 01 leer, página, capítulo, versión, libro, letra, tipografía, perspectiva, comprender, ver, reflexionar,
21114 003 01 libro, letra, Alice Kellen, texto, personaje, historia, revista, experiencia, hobby, Lugares asombros
21112 004 01 rapidez, placer, emoción, felicidad, cultura, información, conocimiento, diversidad, literatura, ensa
21113 005 01 letra, pensar, libro, aprender, información, lenguaje, oración, autor, narración, hoja
21111 006 01 leer, conciencia, entender, información, libro, página, pensar, cuaderno, lápiz, imaginar, destacador
21112 007 01 leer, escribir, libro, lingüística, palabra, frase, cohesión, coherencia
21113 008 01 libro, autor, tarea, manga, poema, letra, cuento, saga, Harry Potter, trabajo, papel, desarrollo, ide
21112 009 01 libro, palabra, Harry Potter, conocimiento, aprendizaje, mundo nuevo, cuento, novela, personaje, fant
21113 010 01 Wattpad, cuento, entretención, aventura, enseñanza, distracción, romance, aprendizaje, historia, conc
21112 011 01 libro, lenguaje, lengua, historia, cuento, leer, padre, niño, leyenda, biblioteca, mito, palabra, sor
noticia, escuela, tarea
21113 012 01 libro, luz, drama, novela, amor, letra, palabra, lámpara, revista, diario, noticia, historia, teatro,
21112 013 01 libro, tiempo, antiguo, novela, reglamento, página, café, lápiz, título, fantasía, estudio, fanfic (t
apunte, capítulo, destacador, manta
21113 014 01 libro, página, portada, lomo, largo, palabra, historia, letra, personaje, tiempo, emoción, hoja, tapa
escritor, editorial, edición
21112 015 01 libro, lentes, concentración, página, papel, novela, autor, fábula, palabra, aprender, imaginación, v
21111 016 01 letra, lentes, palabra, lingüística, ortografía, visión, leer, página, papel, escritura, escribir, ra
soñar, pensar
```

En atención a la metodología detallada en este capítulo, se estudió el léxico disponible de los universitarios de Educación Básica y Letras Hispánica de la Pontificia Universidad Católica de Chile. En el capítulo siguiente se exponen los resultados cuantitativos generales de los datos.

Capítulo 3. Análisis cuantitativos generales

3.1. Consideraciones previas

En este capítulo se exponen los resultados de los análisis cuantitativos generales, entendidos como aquellos que conciernen directamente a las pruebas de disponibilidad léxica, a saber: número de palabras y vocablos, índice cohesión y densidad léxica, promedio de palabras y, sobre todo, el índice de disponibilidad léxica de cada uno de los centros de interés bajo estudio. Entonces, los objetivos que se persiguen son i) examinar las convergencias y divergencias del léxico disponible de estudiantes universitarios de dos carreras distintas, Educación Básica y Letras Hispánicas; y ii) evaluar si el formato digital de recolección de datos se adecua a los principios de DL.

Primeramente, se presentan los análisis de los números de palabras y vocablos; seguidamente, se detallan los resultados del índice de cohesión y densidad léxica; luego, se determinan los promedios de palabras. Se parte de los supuestos de que, por un lado, existen diferencias marcadas en la producción léxica de los encuestados en relación con la carrera que estudian; por otro lado, la riqueza léxica será la misma en grupos distintos a pesar del formato en que se elaboren las pruebas de disponibilidad. Debe destacarse que los análisis generales se han desarrollado a partir de los datos agrupados en los dos subcorpus propuestos, así como por cada una de las muestras por separado. En detalle, el subcorpus 1 está constituido por los listados de Educación Básica (en adelante EB) y Letras Hispánicas (en adelante LH1), recogidos mediante el método tradicional, en formato papel. En tanto que el subcorpus 2, está compuesto por las palabras de Letras Hispánicas recolectadas por el método alternativo, en formato digital (en adelante LH2), y la muestra de LH1. La finalidad de llevar a cabo los análisis generales de esta manera es profundizar en la caracterización cuantitativa del caudal léxico de los grupos.

Después de los análisis generales, se exponen los vocablos más disponibles de cada centro de interés, con el objetivo de comparar los resultados de cada muestra a fin de observar las convergencias y divergencias del caudal léxico de sujetos respecto al programa de estudio y el formato en que contestaron las pruebas.

En las tablas, se resaltan en negritas los cómputos e índices de DL más altos de los centros de interés.

3.2. Análisis del número de palabras

3.2.1. Número total de palabras del subcorpus 1

En la tabla 11 pueden observarse el número total de palabras (NP), los porcentajes que representan y el rango (R) de productividad de cada actualizador del subcorpus 1. El NP es uno de los cómputos generales calculados en los estudios de DL, el cual se refiere al número de casos o piezas léxicas registrados en los listados, independientemente de si se repita o no.

Tabla 11. Número total de palabras por centro de interés del subcorpus 1

Centros de interés	NP	%	Rango
01. La lectura	2668	10,68	6
02. El profesor	2649	10,60	7
03. La educación	2896	11,59	4
04. Juegos y distracciones	2690	10,77	5
05. La escuela	**3335**	**13,35**	**3**
06. Habilidades docentes	2300	9,21	8
07. Partes del cuerpo	**4266**	**17,08**	**1**
08. Comidas y bebidas	**4179**	**16,73**	**2**
Total	24 983		
Promedio	3122,88		

Los 176 informantes del subcorpus 1 reportaron un total de 24 983 palabras en los ocho centros de interés, con promedios de 3122,88 y 141,95 lexías por actualizador y participantes, respectivamente. Las áreas nocionales con los valores más altos –ordenadas de mayor a menor– son: CI07. *Partes del cuerpo*, CI08. *Comidas y bebidas* y CI05. *La escuela*[12], con 4266, 4179 y 3335 piezas léxicas, correspondientemente. En esta misma distribución se organizan dichos CI por rango de productividad, siendo, justamente, el CI07 el que ocupa el primer lugar. Debe acotarse que la suma del NP de estos tres ejes temáticos supera la media aritmética de palabras totales; en efecto, la adición es 11 780 voces. Además, la frecuencia acumulada de estos tres actualizadores alcanza el 60,50 %. Contrariamente, el actualizador menos productivo fue CI06. *Habilidades docentes*[13], con 2300 palabras, ubicándose en el R = 8. Las diferencias entre los CI de los rangos más y menos productivos es 1966 piezas léxicas y 7,87 puntos porcentuales.

En las siguientes subpartes, se detallan los cálculos del NP de cada una de las muestras del subcorpus 1.

[12] Por razones ortotipográficas, el nombre del centro de interés 05. *La escuela: muebles y materiales* se ha simplificado de la siguiente forma: *La escuela*.

[13] Al igual que con el CI05, la identificación del CI06 se sintetizó como sigue: *Habilidades docentes*.

3.2.1.1. Número total de palabras de la muestra 1: Educación Básica en papel

Los estudiantes de la Facultad de Educación escribieron un total de 14 256 palabras, con un promedio de 1782 expresiones por área nocional. El centro de interés más productivo y, por ende, rango 1 es el CI07. *Partes del cuerpo*, con 2452 unidades léxicas, lo que representa el 17,20 % del total general. A este le siguen, descendentemente, los CI08 y CI05, con 2404 (16,86 %) y 1961 (13,76 %) lexías. Asimismo, estos tres ejes temáticos pasan el promedio de palabras –de hecho, únicamente el NP del CI05 es superior– y el porcentaje acumulado es 47,82 %. Al contrario, el eje temático *Habilidades docentes* es el menos productivo, con 1361 lexías, representando el 9,55 % del corpus. Entre este actualizador y el CI07 existe una diferencia de 1091 voces y 7,65 puntos porcentuales. Para más detalles, se presentan los resultados generales de esta muestra en la tabla 12, a continuación.

Tabla 12. Número total de palabras de los estudiantes de Educación Básica

Centros de interés	**TP**	**%**	**Rango**
01. La lectura	1404	9,85	7
02. El profesor	1485	10,42	6
03. La educación	1693	11,88	4
04. Juegos y distracciones	1496	10,49	5
05. La escuela	**1961**	**13,76**	**3**
06. Habilidades docentes	1361	9,55	8
07. Partes del cuerpo	**2452**	**17,20**	**1**
08. Comidas y bebidas	**2404**	**16,86**	**2**
Total	14 256		
Promedio	1782		

3.2.1.2. Número total de palabras de la muestra 2: Letras Hispánicas en papel

Sobre los test de disponibilidad léxica en papel aplicados a los 68 estudiantes de la Facultad de Letras, puede señalarse que el número total de palabras es 10 727, con un promedio por área nocional igual a 1340,88 lexías. Los tres ejes temáticos más productivos y, por consiguiente, con los rangos más altos son: CI07. *Partes del cuerpo*, CI08. *Comidas y bebidas* y CI05. *La escuela*, cuyos valores son: 1814 (16,91 %), 1775 (16,55 %) y 1374 (12,79 %), correspondientemente. Además, los NP de estos CI son los únicos que sobrepasan la media total, logrando, a su vez, alcanzar un porcentaje acumulado de 46,27 %. En línea contraria, el área nocional 06. *Habilidades docentes* es la que ofrece los cómputos más bajos, a saber: NP = 939, lo que representa el 8,75 % del corpus; esto la ubica en el octavo puesto de los rangos de producción. La separación de los CI más y menos productivos es 875 unidades léxicas y 8,16 puntos porcentuales, detalles que se aprecian en la tabla 13.

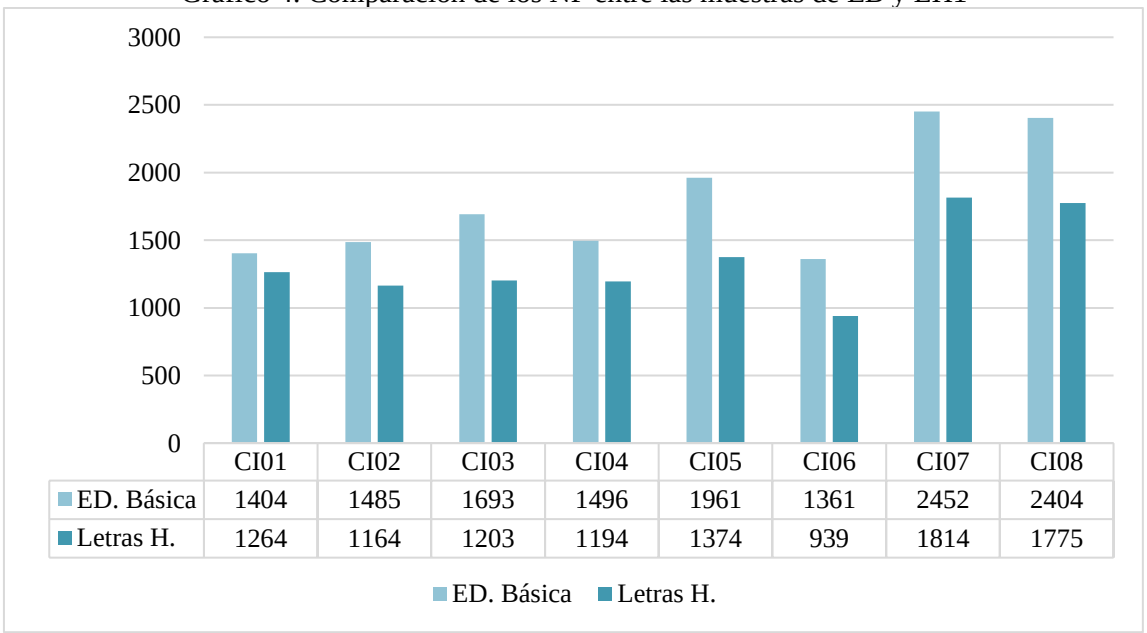
Tabla 13. Número de palabras de los estudiantes de Letras Hispánicas

Centros de interés	TP	%	Rango
01. La lectura	1264	11,78	4
02. El profesor	1164	10,85	7
03. La educación	1203	11,21	5
04. Juegos y distracciones	1194	11,13	6
05. La escuela	**1374**	**12,81**	**3**
06. Habilidades docentes	939	8,75	8
07. Partes del cuerpo	**1814**	**16,91**	**1**
08. Comidas y bebidas	**1775**	**16,55**	**2**
Total	10727		
Promedio	1340,88		

3.2.1.3. Comparación del número de palabras entre las muestras del subcorpus 1

De manera global, los resultados sobre el número de palabras del subcorpus 1 muestran que, primero, los valores de la muestra de Educación Básica superan a los de la de Letras Hispánicas en los ocho centros de interés, alcanzando una diferencia de 3529 piezas léxicas. Esta distinción puede explicarse sobre la base de que el número de encuestados de Educación es mayor al de Letras. Segundo, los CI07, CI08 y CI05, jerárquicamente mencionados, son los más productivos en ambas muestras. Tercero, si bien los CI07, CI08 y CI05 coinciden en los rangos (1.º, 2.º 3.º, respectivamente), se diferencian en que el CI03. *La educación* ocupa el rango 4 en la muestra de EB, mientras que el mismo rango lo tiene el CI01. *La lectura* del conjunto de LH1, como se aprecia en el Gráfico 4.

Gráfico 4. Comparación de los NP entre las muestras de EB y LH1



3.2.2. Número total de palabras del subcorpus 2

En virtud de que el subcorpus 2 está constituido por las respuestas de los dos grupos de Letras Hispánicas –específicamente, por la muestra 2, que contiene los listados de palabras de quienes realizaron las pruebas de DL en papel, y la muestra 3, conformada por las listas elaboradas de manera digital–, se reseñan únicamente el total global del subcorpus y los totales de la muestra 3, ya que los datos de la muestra 2 fueron descritos en el subapartado 3.2.1.2. Sin embargo, sí se presenta el contraste del NP entre ambos grupos.

En el subcorpus 2 se contabilizó un total de 24 268 palabras, con un promedio de 3033,50 lexías por centro de interés. Los actualizadores CI07, CI08 y CI05 –en este orden de mención– ocupan los tres primeros rangos de productividad, ya que sus valores son los más altos, a saber: 4203 (17,32 %), 4145 (17,08 %) y 3061 (12,61 %), respectivamente. En tanto que el CI06 exhibe los cómputos más bajos, NP = 2003 (8,25 %), por lo que se ubica en el último rango. En la Tabla 14, seguidamente, se detallan estos resultados.

Tabla 14. Número total de palabras por centro de interés del subcorpus 2

Centros de interés	**TP**	**%**	**Rango**
01. La lectura	2804	11,55	5
02. El profesor	2560	10,55	6
03. La educación	2738	11,28	5
04. Juegos y distracciones	2754	11,35	4
05. La escuela	**3061**	**12,61**	**3**
06. Habilidades docentes	2003	8,25	7
07. Partes del cuerpo	**4203**	**17,32**	**1**
08. Comidas y bebidas	**4145**	**17,08**	**2**
Total	24 268		
Promedio	3033,50		

3.2.2.1. Número total de palabras de la muestra 3: Letras Hispánicas digital

En la muestra 3, construida a partir de las respuestas de alumnos de Letras Hispánicas que respondieron los test de DL a través de la plataforma electrónica, el número total de palabras equivale a 13 541 lexías, con una media aritmética de 1692,63 unidades por área nocional. Los centros de interés más y menos productivos son *Partes del cuerpo* y *Habilidades docentes*, cuyos NP son 2389 (17,64 %) y 1064 (7,86 %), lo que se traduce en una diferencia de 1325 lexías y 9,78 puntos porcentuales. Además de esto, en el 2.º y 3.er rango se posicionan los CI08 y CI05, cuyos NP son 2370 y 1687, respectivamente. Debe notarse que, en este caso, solamente las áreas nocionales *Partes del cuerpo* y *Comidas y bebidas* sobrepasan el promedio total de respuestas del corpus, y su porcentaje

acumulado suma 35,15 %. En la Tabla 15 se plasman todos los cómputos obtenidos en el subcorpus 2.

Tabla 15. Número total de palabras por área nocional de la muestra 3

Centros de interés	**TP**	**%**	**Rango**
01. La lectura	1540	11,37	5
02. El profesor	1396	10,31	7
03. La educación	1535	11,34	6
04. Juegos y distracciones	1560	11,52	4
05. La escuela	**1687**	**12,46**	**3**
06. Habilidades docentes	1064	7,86	8
07. Partes del cuerpo	**2389**	**17,64**	**1**
08. Comidas y bebidas	**2370**	**17,50**	**2**
Total	13541		
Promedio	1692,63		

3.2.2.2. Comparación del número de palabras entre las muestras del subcorpus 2

Al contrastar el número de palabras de los dos grupos de estudiantes de Letras Hispánicas que componen el subcorpus 2, se observa que el total de piezas léxicas de la muestra recogida de forma digital (NP = 13 541) supera el valor exhibido por la muestra recolectada en papel (NP = 10 727), mostrando una diferencia de 2819 unidades entre ambos grupos. Sobre las áreas nocionales más productivas, tanto en la muestra 2 como en la 3, los CI07, CI08 y CI05 son los que tienen los valores más altos de NP, ocupando los rangos más altos en los dos grupos. Por su parte, el CI06 es el que tiene los resultados más bajos en ambos conjuntos de datos, como se ilustra en la Tabla 16.

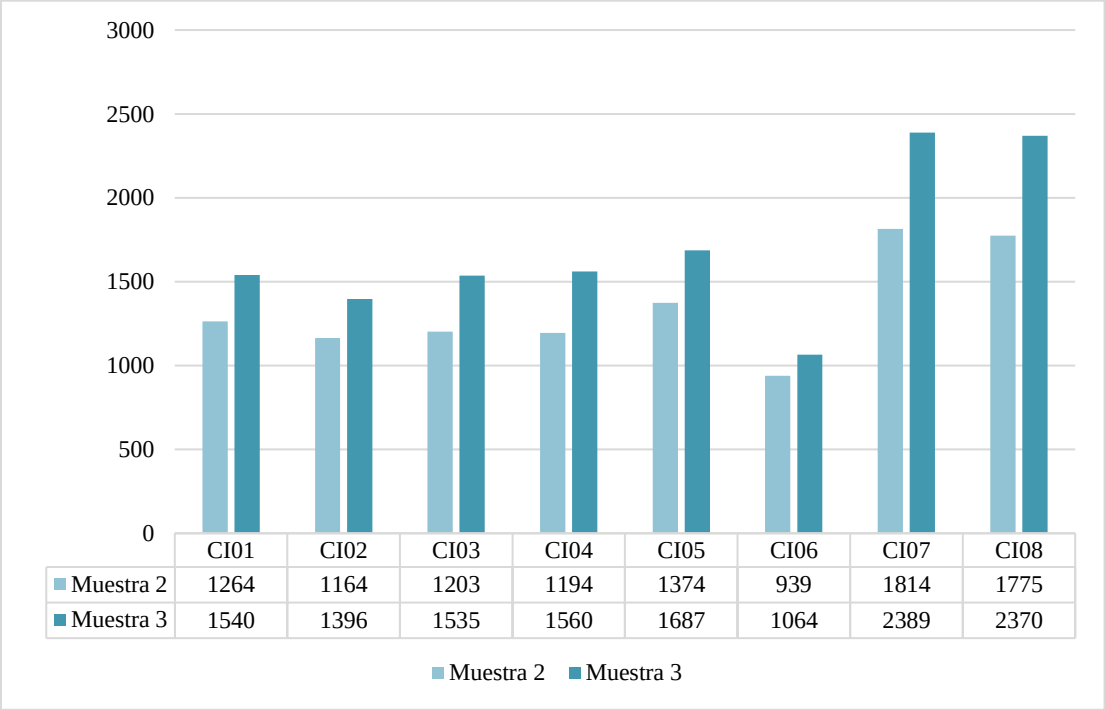
Tabla 16. Comparación de los NP de las dos muestras de Letras Hispánicas

	Muestra digital			**Muestra en papel**		
CI	TP	%	Rango	TP	%	Rango
01. La lectura	1540	11,37	5	**1264**	**11,78**	**4**
02. El profesor	1396	10,31	7	1164	10,85	7
03. La educación	1535	11,34	6	1203	11,21	5
04. Juegos y distracciones	**1560**	**11,52**	**4**	1194	11,13	6
05. La escuela	1687	12,46	3	1374	12,81	3
06. Habilidades docentes	1064	7,86	8	939	8,75	8
07. Partes del cuerpo	2389	17,64	1	1814	16,91	1
08. Comidas y bebidas	2370	17,50	2	1775	16,55	2
Número de palabras	13541			10727		
Promedio de palabras	1692,63			1340,88		

En cuanto a las diferencias por el rango de productividad, se observa que los ejes temáticos CI07, CI08 y CI09 son los que ocupan mayormente los tres primeros rangos, por lo que resulta interesante,

entonces, revisar cuáles son los actualizadores más productivos además de estos tres. En el caso de la muestra tradicional, el área nocional *La lectura* es la que se posiciona en el rango 4. Pero, en la muestra digital, el actualizador que se ubica en el cuarto rango es *Juegos y distracciones*, mientras que el CI01 se halla en el rango 5. En el Gráfico 5 se detallan los resultados del NP en el subcorpus 2.

Gráfico 5. Contraste del NP entre los datos entre LH1 y LH2



	CI01	CI02	CI03	CI04	CI05	CI06	CI07	CI08
Muestra 2	1264	1164	1203	1194	1374	939	1814	1775
Muestra 3	1540	1396	1535	1560	1687	1064	2389	2370

En el Gráfico 5 se aprecian, comparativamente, las diferencias del NP en las dos muestras de Letras Hispánicas por cada centro de interés. Si bien, los valores de la muestra digital son más elevados que los de la tradicional, ambos grupos presentan un patrón distribucional bastante similar, en el que resalta la productividad léxica de las áreas nocionales *Partes del cuerpo* y *Comidas y bebidas*.

3.3. Análisis cuantitativo de vocablos

Una vez conocido y descrito los resultados del número total de palabras del corpus bajo análisis, en esta sección se reseñan los resultados del número de vocablos; es decir, de las unidades léxicas diferentes.

3.3.1. Número de vocablos del subcorpus 1

El total de vocablos aportados por los 176 informantes que conforman el subcorpus 1 es 5837, siendo las áreas nocionales CI03. *La educación* y CI07. *Partes del cuerpo* las que tienen la mayor y menor cantidad de lemas, 867 (14,85 %) y 386 (6,61 %), respectivamente; lo que denota una diferencia de 481 lexías y 8,24 puntos porcentuales entre ambos ejes temáticos. La media aritmética

es 729,63 vocablos por actualizador; este valor es superado por los actualizadores CI03, CI04, CI02, CI06 y CI01, que son los que se ubican en los primeros cinco puestos de rango. En la tabla 17, se detallan los resultados del NV por centros de interés, con sus correspondientes porcentajes y rangos.

Tabla 17. Resultados, generales y por centro de interés, del NV del subcorpus 1

Centros de interés	**NV**	**%**	**Rango**
01. La lectura	781	13,38	5
02. El profesor	823	14,10	3
03. La educación	867	14,85	1
04. Juegos y distracciones	855	14,65	2
05. La escuela	602	10,31	7
06. Habilidades docentes	811	13,89	4
07. Partes del cuerpo	386	6,61	8
08. Comidas y bebidas	712	12,20	6
Total	5837		
Promedio	729,63		

3.3.1.1. Número total de vocablos de la muestra 1: Educación Básica

Tabla 18. Resultados del NV, totales y por CI, de la muestra de EB

Centros de interés	**TV**	**%**	**Rango**
01. La lectura	458	11,61	5
02. El profesor	554	14,04	2
03. La educación	**566**	**14,35**	**1**
04. Juegos y distracciones	546	13,84	3
05. La escuela	435	11,03	6
06. Habilidades docentes	546	13,84	3
07. Partes del cuerpo	299	7,58	7
08. Comidas y bebidas	541	13,71	4
Total	3945		
Promedio	493,13		

En la tabla 18 se presentan los resultados generales y por área nocional del NV de la muestra 1, correspondiente a las respuestas en papel de los estudiantes de Educación Básica. En esta se observa que el total y el promedio general equivalen a 3945 y 493,13 unidades de citas, respectivamente. Los centros de interés más y menos ricos son CI03. *La educación* y CI07. *Partes del cuerpo*, con 566 (14,35 %) y 299 (7,58 %) lemas, correspondientemente. Estos cómputos indican que existe una separación de 267 unidades de cita y 6,77 puntos porcentuales entre estos dos centros de interés. Para los encuestados del área de pedagogía, además del CI03, el CI02. *El profesor* es el segundo léxicamente más ricos, con 554 (14,04 %) lemas. Paralelamente, en la 3.ª posición por rango se

encuentran los CI *Juegos y distracciones* y *Habilidades docentes*, que alcanzaron el mismo NV = 546 (13,84 %).

3.3.1.2. Número total de vocablos de la muestra 2: Letras Hispánicas

En cuanto a los resultados de la muestra 2 –léxico disponible en papel de LH–, el número y promedio son 3638 y 454,75 vocablos, respectivamente. Los actualizadores más y menos ricos son CI03. *La educación* y CI07. *Partes del cuerpo*, con 556 (15,28 %) y 256 (7,04 %), correspondientemente; lo que muestra una separación de 300 unidades de citas y 8,24 puntos porcentuales. En tanto que los centros de interés *Juegos y distracciones*, *La lectura*, *Habilidades docentes* y *El profesor* –ordenados decrecientemente, junto con el CI03 a la cabeza– superan la media aritmética general de vocablos. En la Tabla 19 se detallan los resultados del NV en la muestra 2.

Tabla 19. Resultados del número de vocablos, generales y por CI, de LH1

Centros de interés	TV	%	Rango
01. La lectura	536	14,73	3
02. El profesor	472	12,97	5
03. La educación	556	15,28	1
04. Juegos y distracciones	547	15,04	2
05. La escuela	365	10,03	6
06. Habilidades docentes	478	13,14	4
07. Partes del cuerpo	256	7,04	8
08. Comidas y bebidas	428	11,76	7
Total	3638		
Promedio	454,75		

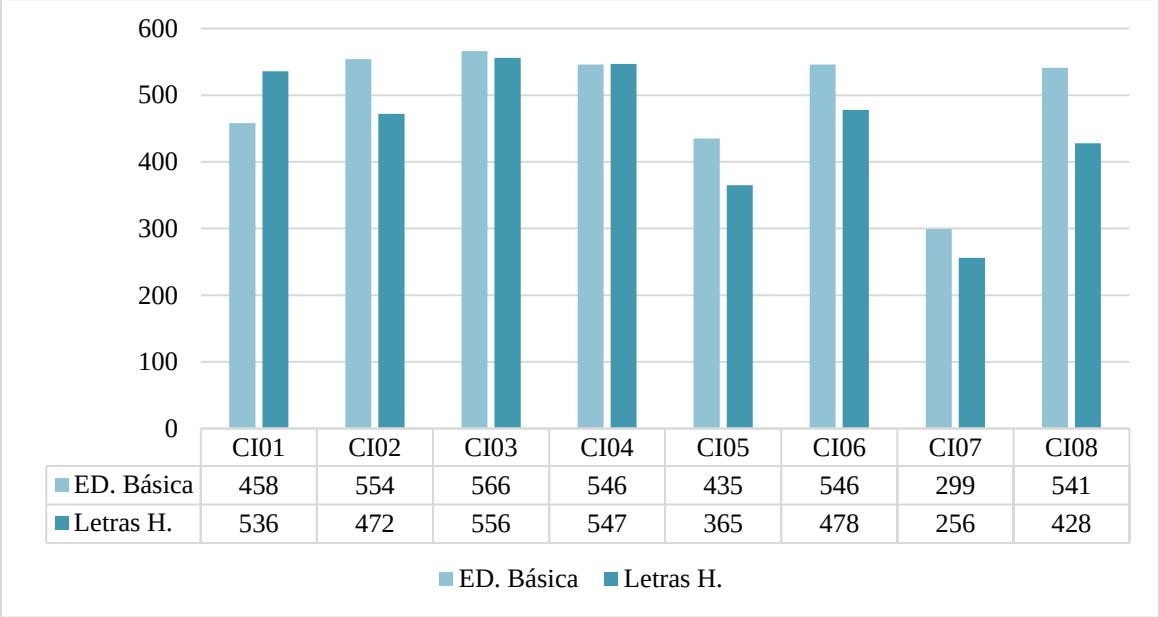
3.3.1.3. Comparación del número de vocablos entre las muestras del subcorpus 1

Al contrastar los resultados del NV de los estudiantes de Educación Básica y los de Letras Hispánicas –que realizaron las pruebas de DL en formato papel o tradicional–, se aprecia que el grupo del área de Pedagogía expone un valor más alto que el de Humanidades, 3945 y 3638 unidades de citas, lo que representa una separación de 307 vocablos. Sobre los actualizadores, los dos grupos convergen en que los centros de interés *La educación* y *Partes del cuerpo* son los que presentan la mayor y menor riqueza de lemas, respectivamente. Si bien los datos indican que existe una coincidencia en que el CI03 ostente el rango 1 en ambas muestras, estas se distinguen en que, por un lado, los CI02 y CI04 ocupan, correspondientemente, los R 2 y 3 en la muestra de EB. Por otro lado, los CI04 y CI01 son los que se ubican en los R 2 y 3 en los datos de LH. Los resultados comparativos del NV del subcorpus 1 pueden verse en la Tabla 20 y en el Gráfico 6.

Tabla 20. Comparación de los resultados del NV entre las muestras del subcorpus 1

CI	Ed. Básica	Rango	Letras H.	Rango
CI01	458	5	536	3
CI02	554	2	472	5
CI03	**566**	**1**	**556**	**1**
CI04	546	3	547	2
CI05	435	6	365	7
CI06	546	3	478	4
CI07	299	7	256	8
CI08	541	4	428	6
Total	3945		3638	

Gráfico 6. Comparación del NV por CI de las muestras del subcorpus 1



3.3.2. Número de vocablos del subcorpus 2

En la tabla 21 pueden apreciarse los resultados del NV del subcorpus 2, compuesto por los datos, recolectados en papel y digitalmente, de estudiantes de Letras Hispánicas. El número total y promedio de vocablos en este corpus son 5766 y 720,75, correspondientemente, siendo el centro de interés *Juegos y distracciones* el que supera en cantidad de unidades de citas a los restantes siete campos nocionales; específicamente, con 945 lemas (16,39 %). Opuestamente, el actualizador *Partes del cuerpo* expone el valor más bajo, a saber: 364 vocablos (6,31 %), lo que indica que existe una separación de 581 unidades léxicas distintas y 10,08 puntos porcentuales entre ambas áreas nocionales.

Tabla 21. Resultados del NV, general y por CI, del subcorpus 2

Centros de interés	NV	%	Rango
01. La lectura	843	14,62	3
02. El profesor	747	12,96	5
03. La educación	875	15,18	2
04. Juegos y distracciones	945	16,39	1
05. La escuela	577	10,01	7
06. Habilidades docentes	750	13,01	4
07. Partes del cuerpo	364	6,31	8
08. Comidas y bebidas	665	11,53	6
Total	5766		
Promedio	720,75		

3.3.2.1. Número de vocablos de Letras Hispánicas en formato digital

Los resultados generales del NV de la muestra 3 –Letras Hispánicas digital– indican que hay un total de 393 vocablos, con una media aritmética de 492,38 unidades de citas por área nocional. Dicho promedio fue superado por los centros de interés –mencionados de mayor a menor rango– *Juegos y distracciones*, *La educación*, *La lectura* y *Comidas y bebidas*, cuyos NV y porcentajes son 649 (16,48 %), 589 (14,95 %), 553 (14,04 %) y 503 (12,77 %) lemas, respectivamente. Por el contrario, el actualizador *Partes del cuerpo* se ubica en el último rango, con 284 (7,21 %) vocablos, lo que indica una distancia de 365 unidades léxicas y 9,27 puntos porcentuales respecto al CI04. En la Tabla 22 se exponen todos los resultados de este cálculo en la muestra digital de LH.

Tabla 22. Resultados del cálculo del NV, general y por CI, de la muestra 3

Centros de interés	TV	%	Rango
01. La lectura	553	14,04	3
02. El profesor	488	12,39	5
03. La educación	589	14,95	2
04. Juegos y distracciones	649	16,48	1
05. La escuela	395	10,03	7
06. Habilidades docentes	478	12,14	6
07. Partes del cuerpo	284	7,21	8
08. Comidas y bebidas	503	12,77	4
Total	3939		
Promedio	492,38		

3.3.2.2. Comparación de los resultados del NV de las muestras del subcorpus 2

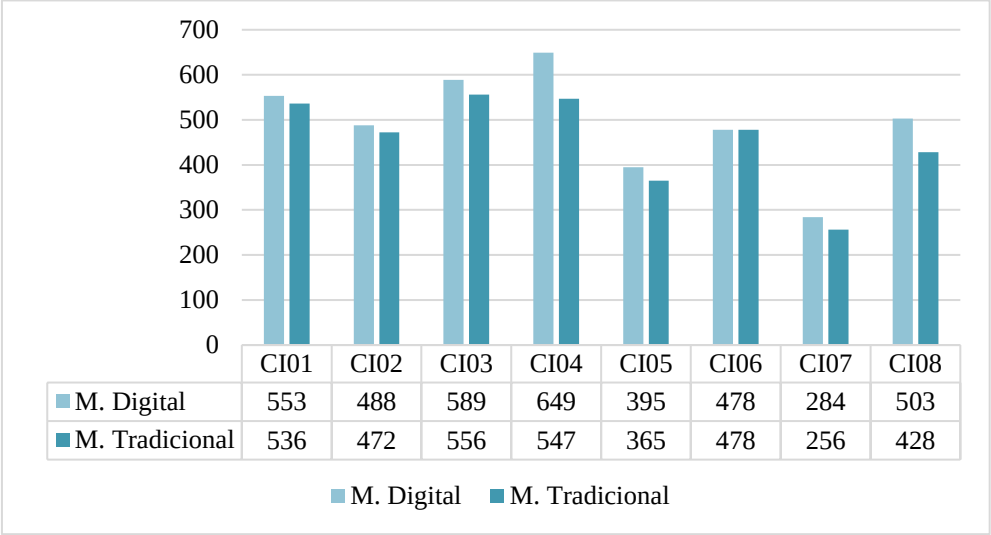
En la tabla 23 puede verse la distribución de los resultados, generales y por actualizador, del número de vocablos de las dos muestras –digital y tradicional– de Letras Hispánicas. La muestra digital suma 3939 vocablos, lo que marca una distinción de 301 unidades de citas más que la muestra

en formato papel (3638). En cuanto a las áreas nocionales, si bien existe una divergencia sobre el CI más rico en lemas entre ambos grupos, las dos muestras coinciden en que los ejes temáticos *La lectura*, *El profesor* y *Partes del cuerpo* se posicionan en los rangos 3, 5 y 8, en el orden dado.

Tabla 23. Comparación de los resultados del NV de las muestras 2 y 3 del subcorpus 2

Centros de interés	Muestra digital			Muestra en papel		
	NV	%	Rango	NV	%	Rango
01. La lectura	553	14,04	3	536	14,73	3
02. El profesor	488	12,39	5	472	12,97	5
03. La educación	589	14,95	2	**556**	**15,28**	**1**
04. Juegos y distracciones	**649**	**16,48**	**1**	547	15,04	2
05. La escuela	395	10,03	7	365	10,03	6
06. Habilidades docentes	478	12,14	6	478	13,14	4
07. Partes del cuerpo	284	7,21	8	256	7,04	8
08. Comidas y bebidas	503	12,77	4	428	11,76	7
Número total de vocablos	3934			3641		

Gráfico 7. Comparación de los resultados del NV por CI de las muestras del subcorpus 2



En síntesis, a pesar de las diferencias relacionadas con la cantidad de vocablos entre las dos muestras, en el Gráfico 7 se observa un patrón bastante similar en la distribución de la riqueza léxica por actualizador en ambos grupos analizados. De hecho, se aprecia poca oscilación del NV, apreciándose una mayor separación en los CI04 y CI08, que exhiben diferencias de 102 y 74 lemas, en dicho orden de mención. Por su parte, el CI06 expone medidas iguales (478 piezas léxicas) en los dos conjuntos de datos. En este sentido, podría deducirse que, indistintamente del formato en el que los sujetos realicen las pruebas de DL, los análisis cuantitativos de los vocablos podrían llegar a dibujar trazos semejantes, siempre que los participantes compartan los mismos saberes sobre los

tópicos o campos semánticos explorados. Sin embargo, esta hipótesis debería ser revisada en otra investigación de DL.

3.4. Análisis de dispersión de palabras

Cuando un informante realiza una prueba de disponibilidad léxica, se activa un conjunto de palabras que pueden estar relacionadas directa o indirectamente con el concepto que evoca el centro de interés. Sin embargo, la actualización del léxico varía de un área nocional a otra. En efecto, algunos CI son robustos en términos de cohesión semántica, ya que las respuestas de los individuos pueden llegar a ser altamente coincidentes, debido a las restricciones del campo semántico, por ejemplo: el cuerpo humano. Opuestamente, otros campos nocionales son débiles, porque los datos varían bastante, exhibiendo una cohesión semántica disgregada (Echeverría *et al.*, 1987; Gómez Devís, 2004). En este marco, dentro de los estudios de DL se recurre a los cálculos de índice de cohesión (IC) y densidad léxica (DL), ya que permiten determinar el grado de homogeneidad léxica de los CI. No obstante, los valores de estos cálculos dependen en gran medida del número de palabras del corpus. Entonces, los cómputos de DL e IC aumentan a medida que asciendan los informantes (Sánchez-Saus, 2011: 263; Hernández Muñoz, 2006: 311). Al respecto, Bartol (2004: 51) afirma que no pueden desarrollarse comparaciones cuantitativas entre estudios que cuenten con diferente número de encuestados. A su vez, el análisis también se ve afectado por los criterios de edición utilizados en las pesquisas. A pesar de estas acotaciones, tanto el IC como la DL son recursos matemáticos probados en la línea de la DL, por lo que en esta tesis doctoral han sido aplicado a los datos, con el objetivo de determinar el grado de compactibilidad u homogeneidad de las palabras en los ocho ejes temáticos aquí analizados.

La densidad léxica es un cómputo que se deduce al dividir el número total de palabras de un CI por el número de vocablos; este ayuda a conocer si el actualizador bajo análisis es compacto o difuso. El grado de coincidencia léxica se determina a partir de cuan alto es el resultado, puesto que, si el valor obtenido se aleja más del 1, mayor será el grado de homogeneidad del eje temático. Por ejemplo, en un grupo de 10 sujetos, si cada uno escriben una palabra diferente en un CI, la DL será 1 (10/10), lo que significa que el vocabulario reportado es difuso y, por lo tanto, el CI es abierto o nada compacto. Por el contrario, si los mismos encuestados escriben la misma palabra en un campo nocional, el valor de DL será 10 (10/1), por lo que se infiere que el CI es homogéneo o cerrado. La DL fue propuesta por Alba en 1995, y ha sido utilizada en diversos trabajos, como los de Carcedo (2001), Hernández Muñoz (2004, 2006), Herranz (2020) y Santos-Díaz (2020), entre otros. Sin embargo, puede señalarse que este cómputo es el menos elaborado de los cálculos utilizados en las

investigaciones de DL (Serfati, 2016: 112), aun así, permite establecer relaciones entre el número de palabras y vocablos de una muestra.

Por su parte, el índice de cohesión (IC) también determina el nivel de coincidencia de las respuestas de los participantes en un centro de interés, a través de un cálculo complejo en el que se combina la división del promedio de palabras por informante y el número de vocablos del actualizador (Echeverría *et al.*, 1987: 68). En este caso, el indicador es un número que va de 0 a 1, donde 1 significa que el eje temático es compacto, ya que traduce que las respuestas de los participantes son 100 % idénticas o coincidentes. En cambio, mientras más alejado de 1 se encuentre el índice, mayor es el nivel de dispersión de los datos. En este caso, se asume que el área nocional explorada es abierta o difusa. A razón de esto, Aabidi (2020: 12) señala que una de las ventajas del IC es que –contrario a la DL– evita la influencia de los sujetos atípicos en la aportación de unidades léxicas del grupo bajo análisis.

Hogaño, no existe un acuerdo sobre cuál debe ser el valor del IC para determinar el grado o nivel de cohesión léxica de un área nocional. Al respecto, Serfati (2020: 113) ha propuesto el cómputo 0,05 de IC como percentil definitorio de compactibilidad de las palabras. Por su parte, Gómez Devís (2004: 121) y Mateus-Ferro y Santiago (2006), con base en el IC, plantean que los centros de interés pueden clasificarse en tres niveles: i) compacto, ii) medianamente compacto y iii) semánticamente difuso, en función de cuanto se acerque o aleje de 1 el valor del índice. No obstante, estos autores no explicitan el rango valórico que debe tener un actualizador para ajustarlo a la taxonomía mencionada. En consonancia con estos argumentos, en esta tesis doctoral los centros de interés han sido catalogados en los tres grados señalados. Sin embargo, con la finalidad de que la clasificación fuese lo más clara posible y, por ende, poder contrastar los resultados de las tres muestras aquí analizadas, se designaron rangos de valor para cada nivel, como se explica a continuación:

- Grado 1: compactibilidad alta, se refiere a aquellos CI cuyos IC se acercan más a 1, específicamente, IC ≥ 0,06.
- Grado 2: compactibilidad media, son los casos en los que los actualizadores tienen IC medios, ubicados en un rango entre 0,03 y 0,05.
- Grado 3: poco cohesionado o difuso, los valores del IC son los menores del conjunto, encontrándose por debajo de 0,03.

3.4.1. Resultados de los índices de dispersión léxica del subcorpus 1

En la Tabla 24, se exponen los resultados de los índices de cohesión y densidad léxica por centro de interés del subcorpus 1. En este conjunto de datos, los centros de interés *Partes del cuerpo*

(IC = 0,0628 y DL = 11,05), *Comidas y bebidas* (IC = 0,0333 y DL = 5,87) y *La escuela* (IC = 0,0315 y DL = 5,54) son los que presentan los cómputos y los rangos de compactibilidad más altos. A su vez, son los únicos que superan las medias aritméticas de IC (0,0273) y DL (4,80). Contrariamente, el actualizador *Habilidades docentes* exhibe los números más bajo (IC = 0,0161 y DL = 2,84), ocupando el rango 8, por lo que se trata de un eje temático difuso.

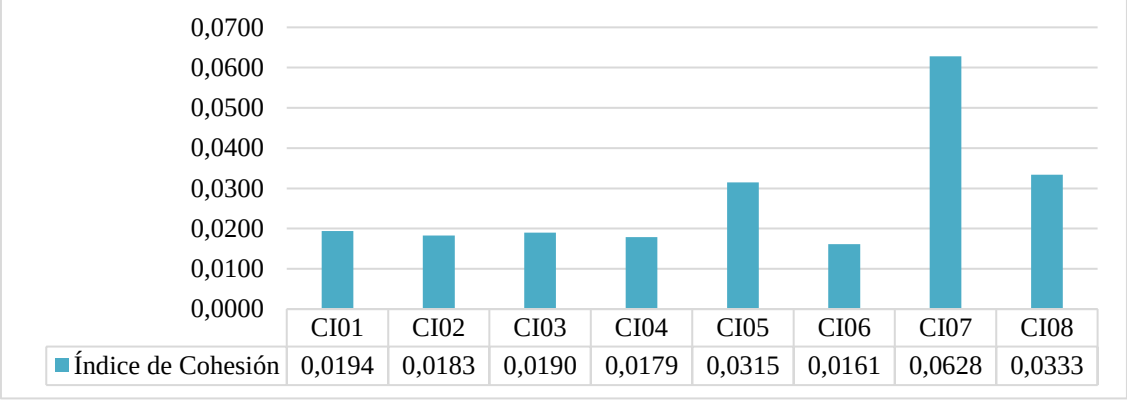
Tabla 24. Índice de Cohesión y Densidad Léxica, de cada CI, del subcorpus 1

Centros de interés	IC	DL	Rango
01. La lectura	0,0194	3,42	4
02. El profesor	0,0183	3,22	6
03. La educación	0,0190	3,34	5
04. Juegos y distracciones	0,0179	3,15	7
05. La escuela	0,0315	5,54	3
06. Habilidades docentes	0,0161	2,84	8
07. Partes del cuerpo	**0,0628**	**11,05**	**1**
08. Comidas y bebidas	0,0333	5,87	2
Promedio	0,0273	4,80	

En cuanto a la agrupación por grado de homogeneidad, se aprecia que solamente un actualizador es compacto, dos son medianamente cerrados y los cinco restantes son difusos, como se observa en el Gráfico 8 y se acota seguidamente:

- Grado 1: *Partes del cuerpo*, con IC = 0,0628 > 0,06.
- Grado 2: *Comidas y bebidas* (0,0333) y *La escuela* (0,0315), con IC > 0,03.
- Grado 3: *La lectura* (0,0194), *La educación* (0,0190), *El profesor* (0,0183), *Juegos y distracciones* (0,0179) y *Habilidades docentes* (0,0161), con IC < 0,03.

Gráfico 8. Resultados generales de los IC por centro de interés del subcorpus 1



3.4.1.1. Índices de dispersión léxica de la muestra de Educación Básica

En la muestra de EB recogida en papel, los actualizadores más cohesionados y, por ende, los que se ubican en los tres primeros rangos, en orden jerárquico, son: *Partes del cuerpo* (IC = 0,0759 y

DL = 8,20), *La escuela* (IC = 0,0417 y DL = 4,51) y *Comidas y bebidas* (IC = 0,0411 y DL = 4,44). Asimismo, estos son los que sobrepasan los promedios del IC y DL. Por su parte, el eje temático más difuso es el sexto, *Habilidades docentes*, con IC = 0,0231 y DL = 2,49. En la tabla 25, se detalla los resultados, por medio de los cuales puede realizarse de la taxonomía de los CI, en relación con su nivel de cohesión, como sigue:

- Nivel 1: *Partes del cuerpo*, con IC = 0,0759 > 0,06.
- Nivel 2: *La escuela* (0,0417) y *Comidas y bebidas* (0,0411), con IC > 0,03.
- Nivel 3: *La lectura* (0,0284), *La educación* (0,0277), *Juegos y distracciones* (0,0254), *El profesor* (0,0248) y *Habilidades docentes* (0,0231), con IC < 0,03.

Tabla 25. Resultados de los IC y DL por CI en la muestra 1: Educación Básica

Centros de interés	IC	DL	Rango
01. La lectura	0,0284	3,07	4
02. El profesor	0,0248	2,68	7
03. La educación	0,0277	2,99	5
04. Juegos y distracciones	0,0254	2,74	6
05. La escuela	0,0417	4,51	2
06. Habilidades docentes	0,0231	2,49	8
07. Partes del cuerpo	0,0759	8,20	1
08. Comidas y bebidas	0,0411	4,44	3
Promedio	0,0360	3,89	

3.4.1.2. Índices de dispersión léxica de la muestra 2: Letras Hispánicas en formato papel

Los resultados de la muestra 2, concerniente a los listados de palabras de los alumnos de Letras Hispánica que realizaron los test de DL de forma tradicional, indican que los centros de interés con los índices y rango de cohesión más altos son *Partes del cuerpo* (IC = 0,1042 y DL = 7,09), *Comidas y bebidas* (IC = 0,0610 y DL = 4,15) y *La escuela* (IC = 0,0554 y DL = 3,76). Además, superan las medias aritméticas del IC y DL. Con base en los datos expuestos en la Tabla 26, los campos nocionales se agrupan por nivel de compactibilidad de la siguiente manera:

- Grado 1: *Parte del cuerpo*, con IC > 0,06.
- Grado 2: *Comidas y bebidas*, *La escuela*, *El profesor* (IC = 0,0363 y DL = 2,47), *La lectura* (IC = 0,0347 y DL = 2,36), *Juegos y distracciones* (IC = 0,0321 y DL = 2,18) y *La educación* (IC = 0,0318 y DL = 2,16).
- Grado 3: *Habilidades y cualidades docentes* (IC = 0,0289 y DL = 1,96).

Tabla 26. Resultados de los IC y DL por CI de LH1

Centros de interés	IC	DL	Rango
01. La lectura	0,0350	2,38	5
02. El profesor	0,0363	2,47	4
03. La educación	0,0318	2,16	7
04. Juegos y distracciones	0,0322	2,19	6
05. La escuela	0,0549	3,74	3
06. Habilidades docentes	0,0289	1,96	8
07. Partes del cuerpo	0,1031	7,01	1
08. Comidas y bebidas	0,0606	4,12	2
Promedio	0,0479	3,25	

3.4.1.3. Contraste de los índices de cohesión entre las muestras del subcorpus 1

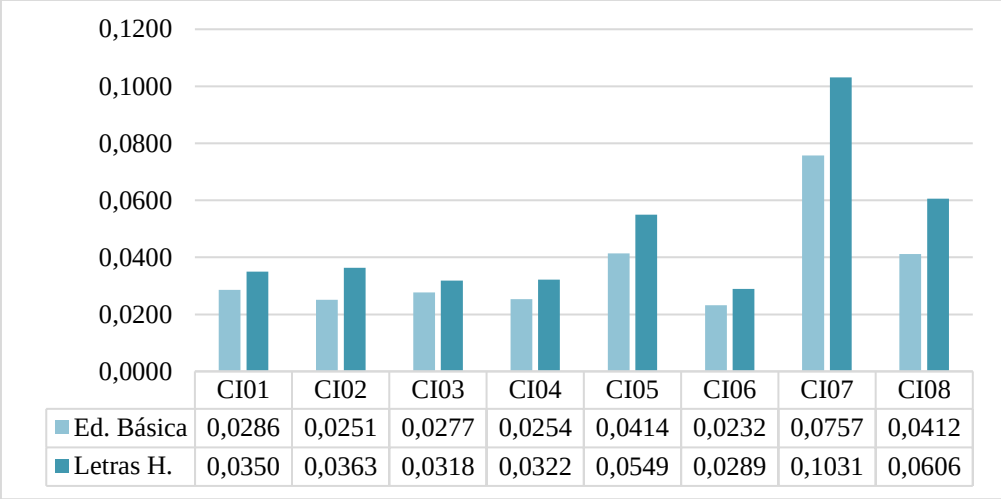
La comparación de los índices de cohesión por área nocional entre los grupos del subcorpus 1 presentan que ambas muestras convergen en que, por un lado, el CI07. *Partes del cuerpo* es el más cerrado y, por lo tanto, compacto de los ocho analizados; mientras que los CI05. *La escuela* y CI08. *Comidas y bebidas* son medianamente cohesionados y cerrados. Por otro lado, el CI06. *Habilidades docentes* es el más abierto y difuso. Sin embargo, las dos muestras difieren en que el conjunto de datos del área de Pedagogía concentró a la mayoría de los actualizadores (*La lectura*, *La educación, Juegos y distracciones*, *El profesor* y *Habilidades docentes*) en el nivel más bajo de cohesión semántica; es decir, resultaron cuantitativamente difusos. En cambio, los datos del área de Humanidades aglutinaron a seis de los ocho actualizadores en el grado de cohesión media (*Comidas y bebidas*, *La escuela*, *El profesor*, *La lectura*, *Juegos y distracciones* y *La educación*), dejando solamente al CI06 en el tercer nivel. Estos análisis se sintetizan en la Tabla 27.

Tabla 27. Distribución de los CI del subcorpus 1, respecto a los grados de compactibilidad

Grados de cohesión	Muestra 1	Muestra 2
Grado o nivel 1	CI07	CI07
Grado o nivel 2	CI05, CI08	CI08, CI05, CI02, CI01, CI04 y CI03
Grado o nivel 3	CI01, CI03, CI04, CI02 y CI06	CI06

El Gráfico 9 complementa los análisis comparativos de los niveles de homogeneidad léxica de los datos de las muestras en formato tradicional de las áreas de Pedagogía y Humanidades. En este se observa que los dos grupos presentan casi el mismo patrón distribucional de los datos léxicos, independientemente de los valores del IC, sean estos altos o bajos, y su ubicación en los niveles de compactibilidad.

Gráfico 9. Contraste de los índices de cohesión léxica por CI de EB y LH1



	CI01	CI02	CI03	CI04	CI05	CI06	CI07	CI08
Ed. Básica	0,0286	0,0251	0,0277	0,0254	0,0414	0,0232	0,0757	0,0412
Letras H.	0,0350	0,0363	0,0318	0,0322	0,0549	0,0289	0,1031	0,0606

3.4.2. Resultados de los índices de dispersión léxica del subcorpus 2

En este subapartado se exponen los resultados de la dispersión léxica, según los índices de cohesión y densidad léxica, de las muestras que constituyen el subcorpus 2 de este estudio. En la Tabla 28 se presentan los IC y DL por centro de interés de todos los datos del subcorpus 2, mientras que en la Tabla 29, se detallan únicamente los resultados de la tercera muestra de la tesis.

Tabla 28. Resultados de los IC y DL por CI del subcorpus 2

Centros de interés	**IC**	**DL**	**Rango**
01. La lectura	0,0213	3,33	5
02. El profesor	0,0220	3,43	4
03. La educación	0,0201	3,13	6
04. Juegos y distracciones	0,0187	2,91	7
05. La escuela	0,0340	5,31	3
06. Habilidades docentes	0,0171	2,67	8
07. Partes del cuerpo	**0,0740**	**11,55**	**1**
08. Comidas y bebidas	0,0400	6,23	2
Promedio	0,0309	4,82	

Tabla 29. Resultados de los IC y DL por CI de la muestra 3

Centros de interés	**IC**	**DL**	**Rango**
01. La lectura	0,0316	2,78	5
02. El profesor	0,0325	2,86	4
03. La educación	0,0296	2,61	6
04. Juegos y distracciones	0,0273	2,40	7
05. La escuela	0,0485	4,27	3
06. Habilidades docentes	0,0253	2,23	8
07. Partes del cuerpo	**0,0956**	**8,41**	**1**
08. Comidas y bebidas	0,0535	4,71	2
Promedio	0,0430	3,78	

Los cálculos de los IC y DL indican que el centro de interés *Partes del cuerpo* es el que ostenta los valores más altos, tanto en los resultados generales del subcorpus 2 (IC = 0,0740 y DL = 11,55) como en los de la muestra recogida de forma digital (IC = 0,0956 y DL = 8,41). Seguidamente, se encuentran los CI08 y CI05, los que, además, junto con el CI07, pasan las medias aritméticas de los IC y DL, generales y de la muestra 3. Al realizar la clasificación por niveles de cohesión semántica de los CI, los valores de los datos totales del subcorpus 2 ubican al CI07, en el primer nivel de compactibilidad; mientras que los CI08 y CI05 se hallan en el segundo nivel; y los restantes, en el tercer nivel. Por su parte, la distribución de los CI de la muestra 3, según los índices de dispersión, es como sigue:

- Grado 1: *Partes del cuerpo*, IC > 0,06.
- Grado 2: *Comidas y bebidas* (IC = 0,0535 y DL = 4,71), *La escuela* (IC = 0,0485 y DL = 4,27), *El profesor* (IC = 0,0325 y DL = 2,86) y *La lectura* (IC = 0,0316 y DL = 2,78), IC > 0,03.
- Grado 3: *La educación* (IC = 0,0296 y DL = 2,61), *Juegos y distracciones* (IC = 0,0273 y DL = 2,40) y *Habilidades docentes* (IC = 0,0253 y DL = 2,23), IC < 0,03.

3.4.2.1. Contraste de los índices de cohesión de las dos muestras del subcorpus 2

La comparación de los resultados de los IC y DL de las dos muestras del subcorpus 2, elaboradas por estudiantes de Letras Hispánicas en formatos papel y digital, indica que existe una alta similitud entre los cómputos de los dos grupos. En primer lugar, en ambos conjuntos de datos, el CI07 es el que ocupa el rango 1 y el grado 1 de cohesión semántica. En segundo lugar, junto con el CI07, los CI08, CI05, CI02, CI01, coinciden en los mismos rangos de compactibilidad léxica en las dos muestras de LH. En tercer lugar, en ambos grupos, el CI06 es el menos cohesionado y, por lo tanto, el más difuso de los ocho, como se aprecia en la Tabla 30.

Tabla 30. Comparación de los IC y DL por CI en las muestras del subcorpus 2

	Muestra digital			**Muestra tradicional**		
Centro de Interés	**IC**	**DL**	**Rango**	**IC**	**DL**	**Rango**
01. La lectura	0,0316	2,78	5	0,0350	2,38	5
02. El profesor	0,0325	2,86	4	0,0363	2,47	4
03. La educación	0,0296	2,61	6	0,0318	2,16	7
04. Juegos y distracciones	0,0273	2,40	7	0,0322	2,19	6
05. La escuela	0,0485	4,27	3	0,0549	3,74	3
06. Habilidades docentes	0,0253	2,23	8	0,0289	1,96	8
07. Partes del cuerpo	**0,0956**	**8,41**	**1**	**0,1031**	**7,01**	**1**
08. Comidas y bebidas	0,0535	4,71	2	0,0606	4,12	2
Promedio	0,0430	3,78		0,0479	3,25	

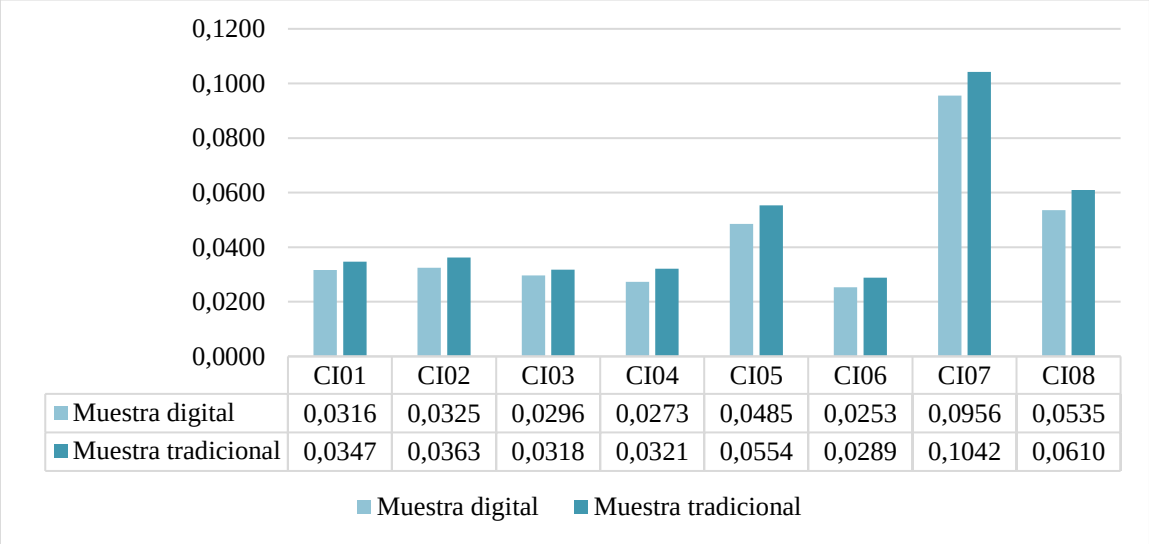
En tanto a la clasificación de los ejes temáticos por niveles de cohesión semántica, los dos grupos convergen en que, por una parte, el centro de interés *Partes del cuerpo* está en el 1.er grado de homogeneidad, lo que significa que es un área nocional cerrada o compacta; a la vez, los actualizadores *Comidas y bebidas*, *La escuela*, *El profesor* y *La lectura* se hallan en el 2.º grado, es decir, son ejes temáticos medianamente compactos. Por otra parte, *Habilidades docentes* está categorizado en el 3.er grado, o sea, es difuso y abierto. Sin embargo, las muestras de LH difieren en que los CI *Juegos y distracciones* y *La educación* están en el segundo nivel, en la muestra recogida en papel; por el contrario, se encuentran en el tercer nivel del conjunto de datos digitales, según se observa en la Tabla 31.

Tabla 31. Distribución de los CI del subcorpus 2 respecto a los grados de compactibilidad

Grados de cohesión	**Muestra 2**	**Muestra 3**
Grado o nivel 1	CI07	CI07
Grado o nivel 2	CI08, CI05, CI02, CI01, CI04 y CI03	CI08, CI05, CI02, CI01
Grado o nivel 3	CI06	CI04, CI03 y CI06

A manera de suplemento de los análisis de los IC, en el Gráfico 10 se observa que los niveles de cohesión por CI trazan un mismo esquema en los dos grupos de Letras Hispánicas, indistintamente del formato en el que se hayan realizado las pruebas de DL. En este se nota claramente la relevancia de los centros de interés 07, 08 y 05, que demuestran el alto nivel de compactibilidad léxica.

Gráfico 10. Contraste de los resultados de los IC por CI entre las muestras del subcorpus 2



	CI01	CI02	CI03	CI04	CI05	CI06	CI07	CI08
Muestra digital	0,0316	0,0325	0,0296	0,0273	0,0485	0,0253	0,0956	0,0535
Muestra tradicional	0,0347	0,0363	0,0318	0,0321	0,0554	0,0289	0,1042	0,0610

En síntesis, como se ha afirmado en estudios previos (cf. Gómez Devís, 2004; Santos Díaz, 2015, 2020; Serfati, 2020), los valores de los IC y DL llegan a depender de la cantidad de informantes y la metodología en la edición de los datos, por lo que resulta un tanto complicado desarrollar un

contraste completamente adecuado entre los resultados obtenidos en corpus diferentes. A pesar de esto, la bibliografía ha demostrado, asimismo, lo fructífero que ha significado la aplicación de los IC y DL en los análisis comparativos, puesto que llegan a mostrar un panorama global acerca de la naturaleza semántica de los centros de interés. En efecto, aunque las tres muestras exploradas en esta tesis cuentan con números dispares de informantes, se ha podido conocer y detallar los grados de homogeneidad léxica de los ocho actualizadores planteados.

3.5. Análisis de los promedios de palabras

En el marco de los análisis cuantitativos generales de DL, el promedio de palabras[14] tiene el objetivo de evaluar la riqueza o productividad léxica, tanto a nivel del grupo de lexicones en general como de cada centro de interés, en particular. Este indicador ha resultado bastante fiable en este tipo de estudio léxico-métrico, puesto que –a diferencia del número de palabras y vocablos; el índice de cohesión y densidad léxica– este no se ve tergiversado o influenciado por factores alterantes como el n muestral, sesgos de respuestas o el nivel de coincidencia de los vocablos. En otras palabras, la media aritmética posee la propiedad matemática de suprimir las distorsiones que pueden llegar a causar las disparidades presentes en las muestras analizadas. Este rasgo convierte al PP en el cálculo ideal para llevar a cabo contrastes intra e intermuestrales de léxico disponible (Gómez Devís, 2004: 114; Hernández Muño, 2006: 309; Santos Díaz, 2020).

Con el fin de optimizar la manera de exponer los datos y realizar un análisis contrastivo más detallado sobre la riqueza léxica de los centros de interés de cada muestra, se ha planteado una categorización de cuatro niveles fundamentada en las medias aritméticas de los actualizadores. Concretamente, cada grupo está definido a partir de la cercanía o lejanía del $\overline{X}$ de cada CI en relación con el promedio total del corpus (cf. Gómez Devís, 2004). La escala de cuatro niveles propuesta se explica a continuación:

- Nivel o grado 1, este reúne a los ejes temáticos que superan la media general por más de tres vocablos; ergo, se refiere a los CI altamente productivos.
- Nivel o grado 2, este aglutina a los centros de interés que están muy cercanos –por arriba o por debajo– al $\overline{X}$ total, con una separación de hasta tres vocablos. Se trata de los CI medianamente productivos.

[14] También denominado *media aritmética* o, simplemente, *media*; se representan con el símbolo $\overline{X}$ y se identifican. Estas, junto con la sigla PP, son las formas utilizadas en esta tesis para referirse al promedio de palabras, a manera de sinónimos.

- Nivel o grado 3, congrega a los actualizadores con tres a cinco vocablos inferiores a la media. Estos son los CI poco productivos.
- Nivel o grado 4, agrupa a las áreas nocionales que se encuentran a más de cinco piezas léxicas por debajo del $\overline{X}$ global del corpus. Se trata de los CI menos productivos.

En los siguientes párrafos, se describen las medias aritméticas de las palabras de los tres grupos bajo análisis. Primero, se muestra una panorámica del PP, global y por CI, del subcorpus. Luego, se presentan los PP, totales y por actualizador, de cada una muestra estudiada. Por último, se comparan las medias aritméticas de cada grupo, identificando la riqueza léxica de cada conjunto de datos.

3.5.1. Promedios de palabras del subcorpus 1

En la Tabla 33 se exhiben los PP, generales y por CI, del subcorpus 1.

Tabla 32. Resultados de los promedios de palabras, generales y por CI, del subcorpus 1

Centros de interés	**PP**	**Rango**
01. La lectura	15,18	6
02. El profesor	15,05	7
03. La educación	16,52	4
04. Juegos y distracciones	15,33	5
05. La escuela	18,94	3
06. Habilidades docentes	13,07	8
07. Partes del cuerpo	24,17	1
08. Comidas y bebidas	23,73	2
Promedio de palabras por informante	141,98	
Media aritmética por CI	17,75	

En este conjunto hay 24 983 unidades léxicas en total, con promedios de 141,95 y 17,74 palabras por participantes y área nocional, respectivamente. Los centros de interés más productivos y, por tanto, con los rangos más altos son *Partes del cuerpo* ($\overline{X}$ = 24,24 y R 1)[15], *Comidas y bebidas* ($\overline{X}$ = 23,74 y R 2) y *La escuela* ($\overline{X}$ = 18,95 y R 3). Asimismo, estos son los únicos que rebasan la media aritmética del grupo ($\overline{X}$ = 17,74). Contrariamente, el eje temático *Habilidades docentes* ostenta el promedio más bajo, a saber: 13,07 palabras, ubicándose en el último rango. Entre los actualizadores más y menos productivos existe una diferencia de 11,17 lexías. En cuanto a los grados de productividad léxica de cada actualizador del subcorpus 1, los datos reflejan que:

- Nivel 1 (altamente productivos): *Partes del cuerpo* y *Comidas y bebidas*, con 6,5 y 6 vocablos, respectivamente, por encima de la media global (17,74).

[15] Como se ha indicado previamente, $\overline{X}$ simboliza el promedio y R el rango de productividad léxica.

- Nivel 2 (medianamente productivo): *La escuela* (1,19 lexías por encima del PP), *La educación* (-1,29)[16], *Juegos y distracciones* (-2,46), *La lectura* (-2,58) y *El profesor* (-2,69).
- Nivel o grado 3 (poco productivo): *Habilidades y cualidades docentes* (-4,67).
- Nivel o grado 4 (menos productivo): ninguno.

3.5.1.1. Resultados de los promedios de palabras por CI de Educación Básica

Como se indicó en el subapartado 3.2.1.1., la muestra creada por los 108 estudiantes del área de Pedagogía está compuesta por 14 256 unidades léxicas. Con base en este resultado, las medias aritméticas por informante y área nocional son 132 y 16,50 lexías, correspondientemente. Los centros de interés con los valores más altos, de mayor a menor, son *Partes del cuerpo* ($\overline{X}$ = 22,70 y R 1), *Comidas y bebidas* ($\overline{X}$ = 22,26 y R 2) y *La escuela* ($\overline{X}$ = 18,16 y R 3), los cuales, a su vez, superaron el promedio general por CI. Hay que destacar que los CI07 y CI08 presentan medias bastante similares, con una separación de, apenas, 0,44 palabras. En cuanto, a la organización de los actualizadores según el grado de productividad, conforme a los resultados detallados en la Tabla 33, se observan la siguiente distribución:

- Grado 1 (altamente productivos): *Partes del cuerpo* y *Comidas y bebidas*, con 6,2 y 5,76 lexías, respectivamente, por sobre la media global (16,50).
- Grado 2 (medianamente productivo): *La escuela* (+1,66), *La educación* (-0,82), *Juegos y distracciones* (-2,65) y *El profesor* (-2,75).
- Grado 3 (poco productivo): *La lectura* (-3,50) y *Habilidades y cualidades docentes* (-3,9).
- Grado 4 (menos productivo): ninguno.

Tabla 33. Promedios de palabras de Educación Básica

Centros de interés	PP	Rango
01. La lectura	13,00	7
02. El profesor	13,75	6
03. La educación	15,68	4
04. Juegos y distracciones	13,85	5
05. La escuela	18,16	3
06. Habilidades docentes	12,60	8
07. Partes del cuerpo	22,70	1
08. Comidas y bebidas	22,26	2
Promedio de palabras por informante	132	
Media aritmética por CI	16,50	

[16] Por razones estilísticas y de edición, entre paréntesis, se indica la cantidad de lexías que distan entre el PP del actualizador y la media general del corpus. Con los signos de sumar (+) y restar (-) se identifican si el número de palabras es superior o inferior, correspondientemente, a la media general.

3.5.1.2. Resultados de los PP por CI de Letras Hispánicas en formato papel

La muestra 2 –fabricada a partir de las respuestas de los 68 estudiantes de Letras Hispánica quienes hicieron los test de DL en papel– está constituida por 10 727 piezas léxicas, con una media aritmética total y por actualizador igual a 157,75 y 19,72 lexías, correlativamente. Sobre la base de los resultados expuestos en la Tabla 34, los actualizadores *Partes del cuerpo* ($\overline{X}$ = 26,68 y R 1), *Comidas y bebidas* ($\overline{X}$ = 26,10 y R 2) y *La escuela* ($\overline{X}$ = 20,21 y R 3) son los más productivos. Contrariamente, el actualizador *Habilidades docentes* ocupa el último rango, con un $\overline{X}$ = 13,81 palabras; de manera que entre los CI más y menos productivo existe una diferencia de 12,87 puntos, un número alto en este contexto. Respecto a los niveles de productividad, los centros de interés se localizan como se describe a continuación:

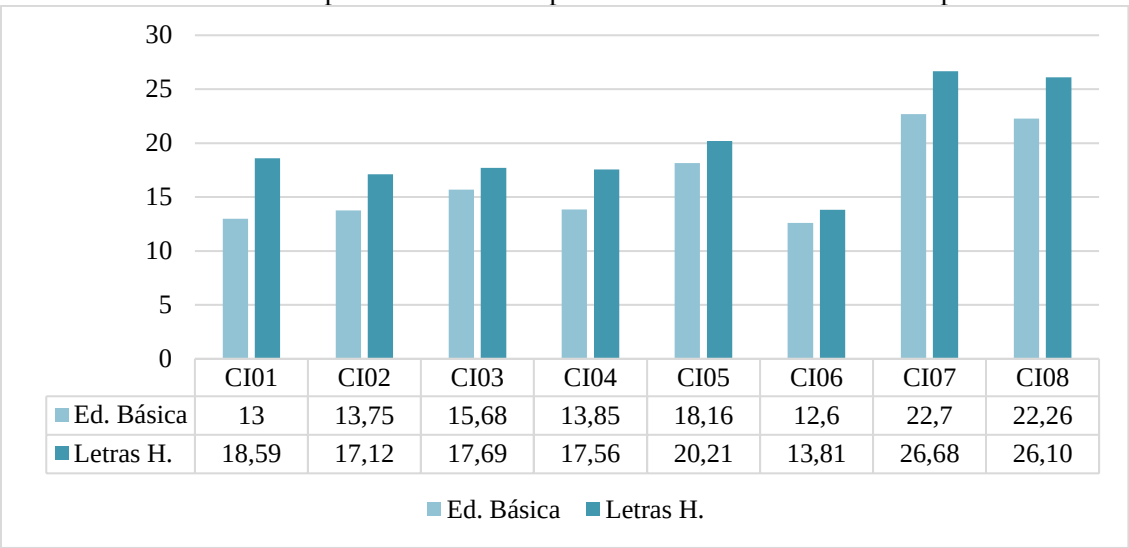
- Grado 1 (altamente productivos): *Partes del cuerpo* (+6,96) y *Comidas y bebidas* (+6,38).
- Grado 2 (medianamente productivo): *La escuela* (+0,49), *La lectura* (-1,13), *La educación* (- 2,03), *Juegos y distracciones* (-2,16) y *El profesor* (-2,60).
- Grado 3 (poco productivo): *Habilidades y cualidades docentes* (-5,91).
- Grado 4 (menos productivo): ninguno.

Tabla 34. Promedio de palabras de los alumnos de LH1

Centros de interés	**PP**	**Rango**
01. La lectura	18,59	4
02. El profesor	17,12	7
03. La educación	17,69	5
04. Juegos y distracciones	17,56	6
05. La escuela	20,21	3
06. Habilidades docentes	13,81	8
07. Partes del cuerpo	26,68	1
08. Comidas y bebidas	26,10	2
Promedio de palabras por informante	157,75	
Media aritmética por CI	19,72	

Con el objetivo de conocer las convergencias y divergencias, en cuanto a la productividad léxica basada en los promedios, de los datos de los estudiantes de Educación Básica y Letras Hispánicas, se han comparado los resultados de ambos grupos, los cuales pueden verse en el Gráfico 11.

Gráfico 11. Comparación de los PP por CI de las muestras del subcorpus 1



	CI01	CI02	CI03	CI04	CI05	CI06	CI07	CI08
Ed. Básica	13	13,75	15,68	13,85	18,16	12,6	22,7	22,26
Letras H.	18,59	17,12	17,69	17,56	20,21	13,81	26,68	26,10

A la luz de estos resultados, pareciera que los alumnos del área de Humanidades tienen una mayor producción de palabras que los del área de Pedagogía, puesto que, en primer lugar, la media aritmética por informante de LH1 (157,75) supera a la de EB (132), denotando una distinción de 25,75 lexías. En segundo lugar, el promedio de unidades léxicas por área nocional de los encuestados de Humanidades (19,72) también está por encima del de los de Pedagogía (16,50), pero la separación en esta medición es de apenas 3,22 expresiones. Por último, los $\overline{X}$ por actualizador de los discentes de Letras son mayores que los de Educación. A pesar de estos resultados, ambos grupos exhiben medias con valores bastante estrechos, llegando a presentar una separación de solamente 1,21 lexía en el CI06, mientras que en los demás CI los PP no llegan a rebasar a 3,98 unidades. No obstante, la mayor diferencia se observa en los PP del eje temático *La lectura*, en cuyo caso la disparidad alcanza un cómputo de 5,59, siendo el grupo de Letras el que refleja la media aritmética más alta ($\overline{X}$ = 18,62 y R 4), mientras que los datos de Educación exhiben un $\overline{X}$ = 13,00 (R 7).

En cuanto a la taxonomía, según los niveles de productividad planteados en esta investigación, en la Tabla 35 se detalla la distribución de los datos, según los cuales se aprecia una mayor confluencia en entre los dos grupos. Pero la disimilitud más grande se observa en la clasificación del CI01 a partir de los resultados del conjunto de Educación Básica, que lleva a ubicar dicho actualizador en el nivel 3, poco productivo, con rango 7. De forma opuesta, el CI01 alcanza el rango 4 y se posiciona en el nivel 2 (medianamente productivo), según los resultados de las medias del grupo de Letras.

Tabla 35. Clasificación de los CI del subcorpus 1 por nivel de productividad

Niveles de productividad	Muestra 1	Muestra 2
Nivel 1	CI07 y CI08	CI07 y CI08
Nivel 2	CI05, CI03, CI04, CI02	CI05, CI01, CI03, CI04 y CI02
Nivel 3	CI01 y CI06	CI06
Nivel 4		

3.5.2. Resultados de los promedios de palabras del subcorpus 2

Tabla 36. Resultados de los PP, generales y por CI, del subcorpus 2

Centros de interés	**PP**	**Rango**
01. La lectura	17,97	4
02. El profesor	16,41	7
03. La educación	17,55	6
04. Juegos y distracciones	17,65	5
05. La escuela	19,62	3
06. Habilidades docentes	12,84	8
07. Partes del cuerpo	26,94	1
08. Comidas y bebidas	26,57	2
Media total por informante	155,56	
Promedio total por CI	19,45	

En la Tabla 36 se detallan los resultados de los PP del subcorpus 2, según esta las medias aritméticas por informante y actualizadores son iguales a 155,56 y 19,45 piezas léxicas, respectivamente. Los PP más altos se observan en los CI *Partes del cuerpo*, *Comidas y bebidas* y *La escuela*, llegando a superar la media total, concretamente, $\overline{X}$ = 26,94 (R 1), $\overline{X}$ = 26,57 (R 2) y $\overline{X}$ = 19,62 (R 3), correlativamente.

3.5.2. Resultados de las medias aritméticas de la muestra 3: Letras Hispánicas en formato digital

El corpus recolectado de forma remota, mediante la página web creada para esta tesis, se compone de 13 541 unidades léxicas, con un promedio total y por centro de interés de 153,88 y 19,23 lexías. Este último valor es sobrepasado únicamente por los actualizadores *Partes del cuerpo* ($\overline{X}$ = 27,15 y R 1) y *Comidas y bebidas* ($\overline{X}$ = 26,93 y R 2). En el lado opuesto se encuentra el CI06. *Habilidades docentes*, con $\overline{X}$ = 12,09 lexías (R 8). Se evidencia que entre el CI07 y CI06 se halla una separación de 15,06 piezas léxicas, es decir, una diferencia bastante notoria. En la tabla 37 se acotan todos los resultados de la muestra 3.

Tabla 37. Promedios de palabras, totales y por CI, de LH2

Centros de interés	PP	Rango
01. La lectura	17,45	5
02. El profesor	15,85	6
03. La educación	17,45	5
04. Juegos y distracciones	17,76	4
05. La escuela	19,17	3
06. Habilidades docentes	12,06	7
07. Partes del cuerpo	27,13	1
08. Comidas y bebidas	26,94	2
Media total por informante	153,82	
Promedio total por CI	19,23	

Al clasificar los ocho centros de interés, en relación con los grados de productividad léxica planteados en esta investigación, se observa la siguiente distribución:

- Nivel 1: *Partes del cuerpo* (+7,92) y *Comidas y bebidas* (+7,7)
- Nivel 2: *La escuela* (-0,06), *Juegos y distracciones* (-1,5), *La lectura* (-1,73) y *La educación* (-1,79).
- Nivel 3: *El profesor*, con -3,37 lexías.
- Nivel 4: *Habilidades y cualidades docentes*, con -7,14 unidades léxicas.

3.5.3. Comparación de las medias entre las muestras de Letras Hispánicas del subcorpus 2.

Tabla 38. Contraste de los PP por CI y rangos en las muestras del subcorpus 2

Centro de Interés	Muestra digital			Muestra tradicional		
	PP	Rango	Nivel	PP	Rango	Nivel
01. La lectura	17,50	5	2	18,59	4	2
02. El profesor	15,86	7	3	17,12	7	2
03. La educación	17,44	6	2	17,69	5	2
04. Juegos y distracciones	17,73	4	2	17,56	6	2
05. La escuela	19,17	3	2	20,21	3	2
06. Habilidades docentes	12,09	8	4	13,81	8	3
07. Partes del cuerpo	**27,15**	**1**	1	**26,68**	**1**	1
08. Comidas y bebidas	26,93	2	1	26,10	2	1
Media total por informante	153,88			157,75		
Promedio total por CI	19,23			19,72		

El contraste de los resultados de las muestras del subcorpus 2 –relativas a los listados escritos por estudiantes de LH, tanto de forma tradicional, en papel, como de manera alternativa, digital– presenta una alta coincidencia en los valores de los promedios de palabras y rangos de los centros de interés en exploración. En efecto, cinco de los ocho ejes temáticos comparten el mismo rango de productividad en cada grupo; puntualmente, de mayor a menor, *Partes del cuerpo* (R1), *Comidas y*

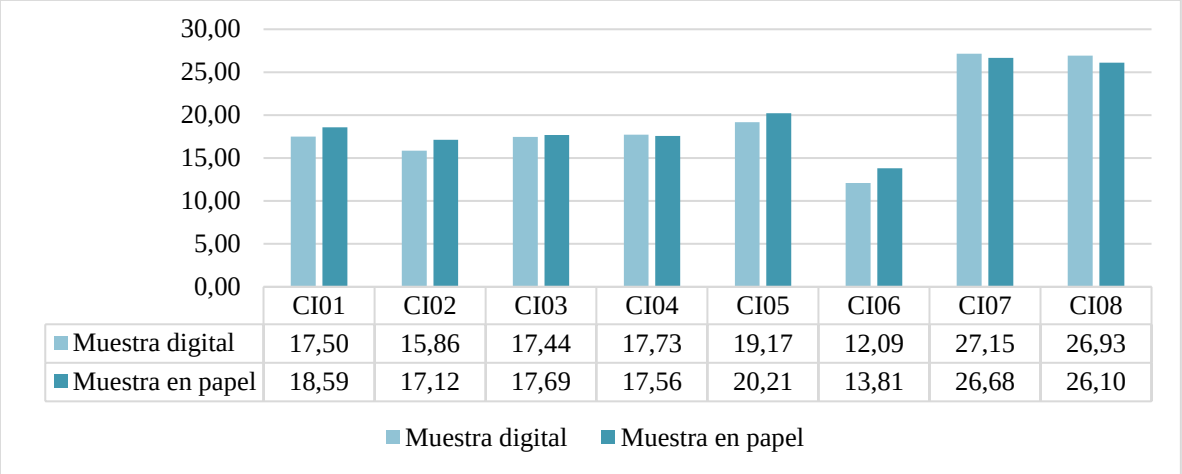
bebidas (R2), *La escuela* (R3), *El profesor* (R7) y *Habilidades docentes* (R8), como se lee en la Tabla 38.

Asimismo, la categorización de las áreas nocionales en relación con el grado de productividad indica que las dos muestras confluyen en agrupar los CI07 y CI08 en el nivel 1; los CI05, CI01, CI04, CI03; pero divergen en que los resultados en papel colocan el CI02 en el segundo grado y el CI06 en el tercero, mientras que los índices del corpus digital ubican el CI02 en el 3.er nivel y el CI06 en el 4.º

En la tabla anterior (38), se aprecia que, si bien los CI07, CI08, CI05, CI03 y CI06 son coincidentes por rango en las dos muestras, se notan algunas divergencias en cuanto al PP de cada área nocional; pero, estas disimilitudes son menores. En primer lugar, las medias de los CI07 y CI08 del conjunto digital están levemente por encima de las del tradicional, diferenciándose apenas por 0,63 y 0,82 lemas en los CI *Partes del cuerpo* y *Comidas y bebidas*, correspondientemente. En segundo lugar, los PP de los CI05, CI03 y CI06 de la muestra en papel están ligeramente por encima de los del corpus alternativo, específicamente: *La escuela* = 0,99; *La educación* = 0,27 y *Habilidades docentes* = 1,75 vocablos de distancia; de hecho, este último expone el contraste más notado. En tercer lugar, el actualizador *La lectura* –que puede considerarse íntimamente ligado a los saberes de los alumnos bajo análisis– ofrece una diferencia de 1,17 vocablos a favor de la muestra tradicional ($\overline{X}$ = 18,62) *versus* ($\overline{X}$ = 17,45) en el corpus alternativo. En cuarto lugar, la media aritmética por eje temático es tenuemente mayor en la muestra tradicional ($\overline{X}$ = 19,71) que en la digital ($\overline{X}$ = 19,23), distinguiéndose apenas por 0,48 puntos. Por último, los estudiantes que escribieron en papel sus respuestas alcanzaron una media por informantes superior a quienes contestaron de manera remota; los primeros reportaron un $\overline{X}$ = 157,68 lexías, mientras que los segundos, un $\overline{X}$ = 163,82, lo que representa una diferencia de 3,86 vocablos entre las dos muestras de LH.

Complementariamente, el Gráfico 12 detalla la baja oscilación de las medias aritméticas por área temática en las dos muestras del subcorpus 2. En efecto, los valores de los PP de cada actualizador no superan siquiera los 1,72 puntos, que se aprecia en el CI06. A su vez, los índices más estrechos se denotan entre los CI04 (0,17), CI03 (0,25), CI07 (0,47) y CI08 (0,83). Entonces, estos datos parecen apuntar –en consideración a los análisis cuantitativos fundados sobre las medias aritméticas– hacia la idea de que el formato digital es tan adecuado como el en papel para la aplicación de los test de disponibilidad léxica.

Gráfico 12. Comparación de los PP de Letras Hispánicas (digital y papel)



3.6. Análisis de los vocablos más disponibles

Tabla 39. Cantidad de vocablos por CI, según algunas propuestas de corte basadas en el IDL

	Ed Básica			Letras H.			Letras H. (digital)		
CI	IDL≥0.1	IDL≥0.2	>25 %	IDL≥0.1	IDL≥0.2	>25 %	IDL≥0.1	IDL≥0.2	>25 %
CI01	11	5	5	17	4	9	18	5	6
CI02	10	2	2	18	6	8	20	4	8
CI03	11	5	9	11	4	8	14	5	8
C04	11	1	4	17	2	7	12	1	3
CI05	19	7	16	24	11	22	20	10	18
CI06	10	4	4	10	5	5	9	4	4
CI07	30	14	28	44	26	32	36	17	34
CI08	27	10	18	36	14	27	41	13	24

En la Tabla 39, se exponen las cantidades hipotéticas de vocablos seleccionados en cada centro de interés, en consideración a tres propuestas de corte, que se cimentan sobre los resultados arrojados por los cálculos de los índices de disponibilidad léxica (IDL). Si bien en los estudios de disponibilidad no existe un criterio único para realizar los cortes y selección del LD de los grupos analizados (Santos Días, 2017c), en esta investigación –con base en un estudio piloto llevado por Martínez-Lara (2021), además de los resultados obtenidos en los trabajos de Carcedo González (2000) y Ávila Muñoz y Sánchez (2010)– se ha considerado realizar el corte en los vocablos con IDL≥0,1. Debe acotarse también que este criterio obedece a que los resultados ofrecen frecuencias, porcentajes y frecuencias acumuladas bajas, los cuales no alcanza ni siquiera el 50 % en algunos casos.

3.6.1. Vocablos más disponibles: comparación intramuestral

En este subapartado se exponen los resultados de los vocablos más disponibles de cada centro de interés de las muestras bajo análisis: Educación Básica, y Letras Hispánicas, en formato tradicional

y digital. Además, se presentan las frecuencias, los porcentajes y las frecuencias acumuladas de los conjuntos de unidades léxicas seleccionadas, según el reporte arrojado por el programa Dispogen.

3.6.1.1. La lectura

Las respuestas recolectadas en la muestra 1 (Facultad de Educación) indica que once vocablos alcanzaron el valor de corte propuesto, siendo el lema *libro* el que ocupa el primer lugar de la lista, con un IDL = 0,7513 y un 81,48 % de aparición; seguido por *leer*, *letra*, *palabra*, *cuento*, *texto*, *autor*, *historia*, *comprensión*, *novela* y *biblioteca*. Debe señalarse que el IDL de *libro* es bastante alto, 5 puntos por encima del segundo vocablo de la lista (*leer*, IDL = 0,2875). En relación con las categorías gramaticales, diez son sustantivos y un verbo (*leer*), como se aprecia en la Tabla 40.

Tabla 40. Vocablos más disponibles del CI *La lectura* de Ed. Básica

no	Vocablos	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	libro	0,7513	88	81,48	0,06268
2	leer	0,2881	37	34,26	0,08903
3	letra	0,2759	39	36,11	0,11681
4	palabra	0,2127	31	28,70	0,13889
5	cuento	0,2007	28	25,93	0,15883
6	texto	0,1691	23	21,30	0,17521
7	autor	0,1452	22	20,37	0,19088
8	historia	0,1330	22	20,37	0,20655
9	comprensión	0,1208	19	17,59	0,22009
10	novela	0,1141	18	16,67	0,23291
11	biblioteca	0,1026	16	14,81	0,24430

Los vocablos con IDL≥0,1 en el grupo de Letras Hispánicas, formato papel, suman dieciséis (15 sustantivos y 1 verbos), siendo *libro* el que ocupa el primer puesto, con un IDL = 0,8798 y 92,65 % de aparición. Las quince unidades léxicas restantes son: *letra*, *leer*, *autor*, *palabra*, *biblioteca*, *página*, *lápiz*, *cuento*, *lector*, *novela*, *literatura*, *historia*, *lentes*, *poesía* y *escritor*. Entre las unidades de citas del 1.er y 2.o rango hay cerca de 5 puntos de diferencia. En la tabla 41, se detallan estos resultados.

Tabla 41. Vocablos más disponibles del CI *La lectura* de LH1

no	Vocablos	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	libro	0,8798	63	92,65	0,04984
2	letra	0,3569	34	50,00	0,07674
3	leer	0,2676	24	35,29	0,09573
4	autor	0,2023	21	30,88	0,11234
5	palabra	0,1977	19	27,94	0,12737
6	biblioteca	0,1817	19	27,94	0,14241
7	página	0,1680	19	27,94	0,15744

8	lápiz	0,1630	18	26,47	0,17168
9	cuento	0,1529	17	25,00	0,18513
10	lector	0,1329	12	17,65	0,19462
11	novela	0,1110	14	20,59	0,20570
12	literatura	0,1077	13	19,12	0,21598
13	historia	0,1037	11	16,18	0,22468
14	lentes	0,1035	10	14,71	0,23259
15	poesía	0,1016	11	16,18	0,24130
16	escritor	0,1011	11	16,18	0,25000

Tabla 42. Vocablos más disponibles del CI *La lectura* de LH2

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	libro	0,8947	81	92,05	0,05260
2	letra	0,3712	40	45,45	0,07857
3	palabra	0,2655	32	36,36	0,09935
4	autor	0,2197	29	32,95	0,11818
5	leer	0,2019	21	23,86	0,13182
6	literatura	0,1693	25	28,41	0,14805
7	novela	0,1519	19	21,59	0,16039
8	cuento	0,1482	20	22,73	0,17338
9	comprensión	0,1454	19	21,59	0,18571
10	biblioteca	0,1432	20	22,73	0,19870
11	página	0,1276	16	18,18	0,20909
12	texto	0,1265	17	19,32	0,22013
13	escritor	0,1158	15	17,05	0,22987
14	escritura	0,1099	14	15,91	0,23896
15	escribir	0,1085	12	13,64	0,24675
16	papel	0,1052	13	14,77	0,25519
17	conocimiento	0,1031	14	15,91	0,26429
18	lápiz	0,1001	13	14,77	0,27273

En la Tabla 42 se muestran los dieciocho vocablos más disponibles de la muestra de Letras Hispánicas recolectada digitalmente. La lista está encabezada por *libro*, cuyo IDL es 0,8947, con 92,05 % de ocurrencias. A este lema le siguen (de mayor a menor): *letra*, *palabra*, *autor*, *leer*, *literatura*, *novela*, *cuento*, *comprensión*, *biblioteca*, *página*, *texto*, *escritor*, *escritura*, *escribir*, *papel*, *conocimiento* y *lápiz*. De estos, 16 son sustantivos y 2 verbos.

3.6.1.2. El profesor

La cantidad de vocablos con IDL≥0,1 es disímil entre las tres muestras analizadas. Específicamente, en los listados léxicos de Educación Básica se contabilizaron diez unidades léxicas

(9 sustantivos y 1 verbo), de las cuales *enseñar* encabeza el vocabulario, con IDL = 0,2293 y 30,56 % de ocurrencias. Seguidamente, se hallan –jerárquicamente– *vocación*, *guía*, *conocimiento*, *educador*, *paciencia*, *colegio*, *enseñanza*, *docente* y *educación*, cuyos índices oscilan muy poco, como se aprecia en la Tabla 43.

Por su parte, el conjunto de Letras Hispánicas, tomado en papel, presenta dieciocho lemas con IDL≥0,1 de los cuales *clase* se encuentra en el rango 1 de disponibilidad, con IDL = 0,2569 y 35,29 % de aparición. Este listado lo integran dieciséis sustantivos y dos verbos, en orden decreciente, según el IDL: *universidad*, *colegio*, *enseñanza*, *conocimiento*, *pizarra*, *alumno*, *prueba*, *enseñar*, *estudio*, *vocación*, *estudiante*, *autoridad*, *aprendizaje*, *aprender*, *docente*, *educación* y *plumón*, según puede observarse en la Tabla 44.

Tabla 43. Vocablos más disponibles del CI *El profesor* de Ed. Básica.

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	enseñar	0,2293	33	30,56	0,02222
2	vocación	0,2118	36	33,33	0,04646
3	guía	0,1637	25	23,15	0,06330
4	conocimiento	0,1499	22	20,37	0,07811
5	educador	0,1459	18	16,67	0,09024
6	paciencia	0,1321	22	20,37	0,10505
7	colegio	0,1319	24	22,22	0,12121
8	enseñanza	0,1190	19	17,59	0,13401
9	docente	0,1088	14	12,96	0,14343
10	educación	0,1047	16	14,81	0,15421

Por último, la muestra digital presenta veinte unidades de citas con el criterio de corte establecido (dieciocho son sustantivos y dos, verbos). El listado lo inicia *enseñanza*, ya que posee el índice más alto, a saber: 0,2690 y 35,23 % de ocurrencias. Los restantes vocablos son: *colegio*, *conocimiento*, *enseñar*, *clase*, *pizarra*, *universidad*, *vocación*, *guía*, *aprendizaje*, *educación*, *estudio*, *libro*, *aprender*, *alumno*, *sala de clase*, *maestro*, *docente*, *materia* y *estudiante*, cuyos valores pueden leerse en la Tabla 45.

Tabla 44. Vocablos más disponibles del CI *El profesor* de LH1

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	clase	0,2569	24	35,29	0,02062
2	universidad	0,2556	28	41,18	0,04467
3	colegio	0,2353	23	33,82	0,06443
4	enseñanza	0,2163	18	26,47	0,07990
5	conocimiento	0,2159	21	30,88	0,09794
6	pizarra	0,2091	22	32,35	0,11684
7	alumno	0,1854	18	26,47	0,13230

8	prueba	0,1769	20	29,41	0,14948
9	enseñar	0,1731	14	20,59	0,16151
10	estudio	0,1490	14	20,59	0,17354
11	vocación	0,1448	16	23,53	0,18729
12	estudiante	0,1283	13	19,12	0,19845
13	autoridad	0,1256	11	16,18	0,20790
14	aprendizaje	0,1239	13	19,12	0,21907
15	aprender	0,1208	11	16,18	0,22852
16	docente	0,1165	9	13,24	0,23625
17	educación	0,1155	13	19,12	0,24742
18	plumón	0,1026	12	17,65	0,25773

Tabla 45. Vocablos más disponibles del CI *El profesor* de LH2

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	enseñanza	0,2690	31	35,23	0,02221
2	colegio	0,2605	35	39,77	0,04728
3	conocimiento	0,2405	32	36,36	0,07020
4	enseñar	0,2168	24	27,27	0,08739
5	clase	0,1956	23	26,14	0,10387
6	pizarra	0,1752	24	27,27	0,12106
7	universidad	0,1691	24	27,27	0,13825
8	vocación	0,1664	22	25,00	0,15401
9	guía	0,1648	18	20,45	0,16691
10	aprendizaje	0,1502	18	20,45	0,17980
11	educación	0,1475	19	21,59	0,19341
12	estudio	0,1399	17	19,32	0,20559
13	libro	0,1286	20	22,73	0,21991
14	aprender	0,1255	17	19,32	0,23209
15	alumno	0,1230	16	18,18	0,24355
16	sala de clase	0,1199	19	21,59	0,25716
17	maestro	0,1127	12	13,64	0,26576
18	docente	0,1073	10	11,36	0,27292
19	materia	0,1057	16	18,18	0,28438
20	estudiante	0,1054	12	13,64	0,29298

3.6.1.3. La educación

Las muestras recolectadas en papel de Ed. Básica y Letras Hispánicas coinciden en que los vocablos *profesor* y *colegios* son los más disponibles de ambas listas, con rangos de disponibilidad 1 y 2, correspondientemente. La muestra de la Facultad de Educación presenta doce lemas con IDL≥0,1 –nueve sustantivos y tres verbos–, a saber: *aprendizaje*, *aprender*, *enseñar*, *alumno, conocimiento, universidad*, *libro*, *niño*, *leer* y *matemáticas*. Por su parte, el grupo de Letras exhibe once unidades de citas con el valor requerido, diez son sustantivos y un verbo: *universidad, aprender, alumno,*

conocimiento, *calidad*, *aprendizaje*, *estudiante*, *libro*, *derecho* a los que deben sumarse los dos arriba mencionados. Los detalles de estos resultados pueden leerse en las Tablas 46 y 47.

Tabla 46. Vocablos más disponibles del CI *La educación* de Ed. Básica.

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	profesor	0,4642	66	61,11	0,03898
2	colegio	0,2883	49	45,37	0,06793
3	aprendizaje	0,2483	40	37,04	0,09155
4	aprender	0,2266	37	34,26	0,11341
5	enseñar	0,2042	30	27,78	0,13113
6	alumno	0,1908	29	26,85	0,14826
7	conocimiento	0,1722	27	25,00	0,16421
8	universidad	0,1672	34	31,48	0,18429
9	libro	0,1478	27	25,00	0,20024
10	niño	0,1458	26	24,07	0,21559
11	leer	0,1083	19	17,59	0,22682
12	matemáticas	0,1037	18	16,67	0,23745

Tabla 47. Vocablos más disponibles del CI *La educación* de LH1.

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	profesor	0,4334	41	60,29	0,03408
2	colegio	0,3656	33	48,53	0,06151
3	universidad	0,2621	28	41,18	0,08479
4	aprender	0,2173	21	30,88	0,10224
5	alumno	0,1830	18	26,47	0,11721
6	conocimiento	0,1615	18	26,47	0,13217
7	calidad	0,1565	14	20,59	0,14381
8	aprendizaje	0,1331	13	19,12	0,15461
9	estudiante	0,1314	15	22,06	0,16708
10	libro	0,1300	17	25,00	0,18121
11	derecho	0,1117	11	16,18	0,19036

Al igual que en las muestras 1 y 2, en formato papel, la palabra más disponible en el diccionario de los datos de Letras Hispánicas recogido digitalmente es *profesor*, con un IDL = 0,4775 y 64,77 % de ocurrencia. En este corpus las unidades de citas con IDL≥0,1 suman catorce, trece sustantivos y un verbo. Estas son (organizadas por rango de mayor a menor): *conocimiento*, *colegio*, *aprendizaje*, *alumno*, *libro*, *aprender*, *universidad*, *derecho*, *necesario*, *estudiante*, *importante*, *enseñanza* y *cuaderno*, como se acota en la Tabla 48.

Tabla 48. Vocablos más disponibles del CI *La educación* de LH2

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	profesor	0,4775	57	64,77	0,03713
2	conocimiento	0,2424	31	35,23	0,05733

3	colegio	0,2242	32	36,36	0,07818
4	aprendizaje	0,2153	27	30,68	0,09577
5	alumno	0,2129	27	30,68	0,11336
6	libro	0,1872	26	29,55	0,13029
7	aprender	0,1626	21	23,86	0,14397
8	universidad	0,1614	26	29,55	0,16091
9	derecho	0,1601	18	20,45	0,17264
10	necesario	0,1467	16	18,18	0,18306
11	estudiante	0,1411	19	21,59	0,19544
12	importante	0,1370	16	18,18	0,20586
13	enseñanza	0,1143	14	15,91	0,21498
14	cuaderno	0,1111	17	19,32	0,22606

3.6.1.4. Juegos y distracciones

El cuarto eje temático analizado es *Juegos y distracciones*. En la muestra 1 (Educación Básica) se aprecian once vocablos con valores de IDL≥0,1, todos sustantivos, siendo el más disponible *diversión* (IDL = 0,2004 y 26,85 %). Por su parte, en la muestra 2 (Letras Hispánicas, en papel), se observan diecisiete casos, de los cuales quince son sustantivos y dos, verbos. La palabra más disponible de este grupo es *compu(tador/a)* (IDL = 0,2757/ 39,71 %). Por último, en la muestra recogida con la página web *ad hoc*, el número de unidades léxicas con IDL≥0,1 suman doce, once sustantivos y un verbo, siendo *amigo* la que encabeza la lista (IDL = 0,2115 y 37,50 %). Estos resultados se detallan en las Tablas 49, 50 y 51.

Tabla 49. Vocablos más disponibles del CI *Juegos y distracciones* de Ed. Básica.

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	diversión	0,2004	29	26,85	0,0193
2	entretención	0,1687	24	22,22	0,0354
3	juego de mesa	0,1602	26	24,07	0,0527
4	celular	0,1521	27	25,00	0,0707
5	amigo	0,1431	30	27,78	0,0907
6	Monopoly	0,1342	20	18,52	0,1041
7	niño	0,1267	19	17,59	0,1167
8	película	0,1080	19	17,59	0,1294
9	tiempo libre	0,1009	16	14,81	0,1401
10	compu(tador/a)	0,1008	18	16,67	0,1521
11	videojuego	0,1006	13	12,04	0,1608

Tabla 50. Vocablos más disponibles del CI *Juegos y distracciones* de LH1.

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	compu(tador/a)	0,2757	27	39,71	0,0225
2	carta	0,2052	19	27,94	0,0384
3	leer	0,1990	18	26,47	0,0534
4	celular	0,1942	18	26,47	0,0684
5	película	0,1854	19	27,94	0,0842
6	ocio	0,1839	17	25,00	0,0984
7	música	0,1796	21	30,88	0,1159
8	videojuego	0,1789	15	22,06	0,1284
9	tiempo libre	0,1598	15	22,06	0,1410
10	tele(visión)	0,1349	14	20,59	0,1526
11	serie	0,1332	14	20,59	0,1643
12	diversión	0,1300	12	17,65	0,1743
13	internet	0,1297	12	17,65	0,1843
14	libro	0,1281	13	19,12	0,1952
15	amigo	0,1141	14	20,59	0,2068
16	dormir	0,1114	14	20,59	0,2185
17	juego de mesa	0,1104	10	14,71	0,2269

Tabla 51. Vocablos más disponibles del CI Juegos y distracciones de LH2

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	amigo	0,2115	33	37,50	0,0211
2	diversión	0,1962	24	27,27	0,0365
3	videojuego	0,1688	20	22,73	0,0493
4	película	0,1496	23	26,14	0,0640
5	leer	0,1383	17	19,32	0,0749
6	entretención	0,1270	15	17,05	0,0845
7	compu(tador/a)	0,1247	17	19,32	0,0953
8	internet	0,1233	14	15,91	0,1043
9	tiempo libre	0,1171	16	18,18	0,1145
10	música	0,1142	17	19,32	0,1254
11	serie	0,1133	18	20,45	0,1369
12	celular	0,1098	14	15,91	0,1459

3.6.1.5. La escuela: muebles y materiales

El centro de interés *La escuela: muebles y materiales* es uno de los ocho actualizadores bajo análisis, y es el que más vocablos con IDL≥0,1 ostenta en las tres muestras exploradas. En los datos de Educación Básica se hallan diecinueve lexías con dicha característica, mientras que en los de Letras Hispánicas (método en papel) se aprecian veinticuatro y, por último, en la muestra digital se cuentan veinte lemas. Debe resaltarse que, en los tres conjuntos de datos, los vocablos más disponibles son

mesa y *silla*, los cuales ocupan el primer y segundo rango, respectivamente, en los tres corpus. En las Tablas 52, 53 y 54 pueden apreciarse con más detalles los resultados obtenidos en esta área nocional.

Tabla 52. Vocablos más disponibles del CI *La escuela* de Ed. Básica

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	mesa	0,6497	89	82,41	0,0454
2	silla	0,6225	90	83,33	0,0912
3	pizarra	0,5298	87	80,56	0,1356
4	lápiz	0,3864	64	59,26	0,1682
5	plumón	0,3598	65	60,19	0,2013
6	cuaderno	0,2549	46	42,59	0,2248
7	estuche	0,2144	40	37,04	0,2452
8	libro	0,1993	45	41,67	0,2681
9	escritorio	0,1882	31	28,70	0,2839
10	sala (de clase)	0,1859	33	30,56	0,3007
11	cartulina	0,1848	38	35,19	0,3201
12	ventana	0,1807	45	41,67	0,3430
13	goma (de borrar)	0,1601	39	36,11	0,3629
14	patio	0,1472	30	27,78	0,3782
15	compu(tador/a)	0,1284	33	30,56	0,3950
16	proyector	0,1257	29	26,85	0,4098
17	borrador	0,1216	26	24,07	0,4230
18	estante	0,1207	26	24,07	0,4363
19	tijera	0,1065	25	23,15	0,4490

Tabla 53. Vocablos más disponibles del CI *La escuela* de LH1

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	mesa	0,7406	59	86,76	0,0430
2	silla	0,7108	58	85,29	0,0853
3	lápiz	0,6244	58	85,29	0,1276
4	pizarra	0,5776	52	76,47	0,1656
5	cuaderno	0,4717	51	75,00	0,2028
6	plumón	0,4077	43	63,24	0,2341
7	estuche	0,3545	38	55,88	0,2619
8	goma (de borrar)	0,3308	37	54,41	0,2888
9	libro	0,3002	35	51,47	0,3144
10	borrador	0,2373	28	41,18	0,3348
11	proyector	0,2070	23	33,82	0,3516
12	regla	0,1912	22	32,35	0,3676
13	mochila	0,1835	25	36,76	0,3858
14	compu(tador/a)	0,1802	22	32,35	0,4019
15	escritorio	0,1794	19	27,94	0,4158
16	pegamento	0,1670	22	32,35	0,4318

17	tijera	0,1651	21	30,88	0,4471
18	sala (de clase)	0,1486	17	25,00	0,4595
19	destacador	0,1394	16	23,53	0,4712
20	corrector	0,1358	17	25,00	0,4836
21	cartulina	0,1338	17	25,00	0,4960
22	hoja	0,1318	18	26,47	0,5091
23	ventana	0,1112	15	22,06	0,5201
24	casillero	0,1014	11	16,18	0,5281

Tabla 54. Vocablos más disponibles del CI *La escuela* de LH2

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	mesa	0,6837	73	82,95	0,0433
2	silla	0,6581	76	86,36	0,0883
3	lápiz	0,6000	71	80,68	0,1304
4	pizarra	0,5007	65	73,86	0,1689
5	cuaderno	0,4326	59	67,05	0,2039
6	libro	0,3633	56	63,64	0,2371
7	plumón	0,3305	51	57,95	0,2673
8	estuche	0,2641	41	46,59	0,2916
9	escritorio	0,2391	28	31,82	0,3082
10	goma (de borrar)	0,2341	37	42,05	0,3302
11	regla	0,1734	29	32,95	0,3474
12	estante	0,1715	32	36,36	0,3663
13	borrador	0,1702	31	35,23	0,3847
14	tijera	0,1490	27	30,68	0,4007
15	destacador	0,1418	25	28,41	0,4155
16	corrector	0,1369	26	29,55	0,4309
17	mochila	0,1364	28	31,82	0,4475
18	compu(tador/a)	0,1328	24	27,27	0,4618
19	cartulina	0,1228	20	22,73	0,4736
20	pupitre	0,1163	13	14,77	0,4813

3.6.1.6. Habilidades y cualidades docentes

Habilidades y cualidades docentes es el sexto centro de interés analizado, además, como se ha mencionado, es uno de los novedosos en los estudios de DL. Conocer el léxico disponible de este campo resulta interesante, puesto que mapea un área nocional que no había sido descrita y permite recopilar datos léxicos acerca de las nociones y representaciones que los estudiantes de las áreas de Pedagogía y Humanidades tienen y evocan sobre los conceptos de las habilidades y las cualidades de los docentes. En este contexto, las dos muestras que componen el subcorpus 1 reflejan nueve vocablos con IDL≥0,1, mientras que la tercera muestra exhibe solo nueve. Tanto en los listados de Educación

Básica (Tabla 55) como en los de Letras Hispánica, recolectados *online*, (Tabla 57) se observa que la palabra más disponible es *paciencia*, cuyos IDL son 0,3319 y 0,4038, respectivamente. En cambio, la más disponible de los datos de Letras H., en papel, (Tabla 56) es *empatía* (IDL = 0,2884 y 39,71 %).

Tabla 55. Vocablos más disponibles del CI *Habilidades docentes* de EB

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	paciencia	0,3319	43	39,81	0,0316
2	empatía	0,2885	40	37,04	0,0610
3	enseñar	0,2338	34	31,48	0,0860
4	vocación	0,2049	29	26,85	0,1073
5	escuchar	0,1723	22	20,37	0,1234
6	comprensión	0,1666	23	21,30	0,1403
7	respeto	0,1584	24	22,22	0,1580
8	conocimiento	0,1251	19	17,59	0,1719
9	aprender	0,1189	17	15,74	0,1844
10	responsabilidad	0,1139	18	16,67	0,1976

Tabla 56. Vocablos más disponibles del CI *Habilidades docentes* de LH1

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	empatía	0,2884	27	39,71	0,0288
2	comprensión	0,2566	24	35,29	0,0543
3	vocación	0,2566	23	33,82	0,0788
4	paciencia	0,2539	22	32,35	0,1022
5	enseñar	0,2193	17	25,00	0,1203
6	amabilidad	0,1553	15	22,06	0,1363
7	conocimiento	0,1552	14	20,59	0,1512
8	respeto	0,1368	13	19,12	0,1651
9	explicar	0,1226	11	16,18	0,1768
10	sabiduría	0,1081	11	16,18	0,1885

Tabla 57. Vocablos más disponibles del CI *Habilidades docentes* de LH2

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	paciencia	0,4038	45	51,14	0,0424
2	empatía	0,3910	40	45,45	0,0801
3	comprensión	0,2140	24	27,27	0,1027
4	vocación	0,2018	23	26,14	0,1244
5	enseñar	0,1224	14	15,91	0,1376
6	responsabilidad	0,1190	14	15,91	0,1508
7	escuchar	0,1130	15	17,05	0,1649
8	amabilidad	0,1089	14	15,91	0,1781
9	respeto	0,1064	15	17,05	0,1923

3.6.1.7. Partes del cuerpo

El séptimo eje temático bajo análisis, *Parte del cuerpo*, es uno de los correspondientes a los propuesto por Gougenheim *et al.* (1964) y también por el PPHLD; en otras palabras, conforma los denominados centros de interés tradicionales. En estudios previos, este ha resultado muy productivo, lo cual puede confirmarse con los datos de esta tesis, puesto que ha sido el actualizador con el número más alto de vocablos con IDL≥0,1 en las tres muestras aquí analizadas. En efecto, la muestra 1, Educación Básica (Tabla 58), exhibe treinta vocablos con el cómputo de corte, mientras que el grupo de Letras Hispánicas, formato físico, (Tabla 59) suma cuarenta y cuatro; y el conjunto de datos de la muestra digital (Tabla 60), treinta y seis unidades de citas. En el primero grupo explorado, la palabra más disponible es *mano* (IDL = 0,5983 y 80,56 %). En tanto que, en las dos muestras de Letras, en papel y digital, el vocablo más disponible es *ojo*, cuyos respectivos IDL son 0,7183 y 0,6763.

Tabla 58. Vocablos más disponibles del CI *Partes del cuerpo* de EB.

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	mano	0,5983	87	80,56	0,0355
2	ojo	0,5958	92	85,19	0,0730
3	cabeza	0,5208	72	66,67	0,1024
4	pierna	0,5139	88	81,48	0,1383
5	brazo	0,5072	79	73,15	0,1705
6	pie	0,4772	87	80,56	0,2060
7	dedo	0,4227	79	73,15	0,2382
8	nariz	0,3649	73	67,59	0,2679
9	uña	0,3485	73	67,59	0,2977
10	oreja	0,3418	71	65,74	0,3267
11	boca	0,3366	64	59,26	0,3528
12	pelo	0,3035	54	50,00	0,3748
13	corazón	0,2516	49	45,37	0,3948
14	rodilla	0,2318	51	47,22	0,4156
15	cuello	0,1838	41	37,96	0,4323
16	cerebro	0,1579	33	30,56	0,4458
17	diente	0,1530	38	35,19	0,4613
18	espalda	0,1472	36	33,33	0,4759
19	codo	0,1465	39	36,11	0,4918
20	pene	0,1438	28	25,93	0,5033
21	estómago	0,1372	40	37,04	0,5196
22	lengua	0,1363	34	31,48	0,5334
23	hombro	0,1307	28	25,93	0,5449
24	cara	0,1298	19	17,59	0,5526
25	pulmón	0,1292	37	34,26	0,5677
26	muslo	0,1217	32	29,63	0,5808

27	ceja	0,1206	27	25,00	0,5918
28	tobillo	0,1088	27	25,00	0,6028
29	pecho	0,1071	25	23,15	0,6130
30	vagina	0,1007	22	20,37	0,6219

Tabla 59. Vocablos más disponibles del CI *Partes del cuerpo* de LH1

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	ojo	0,7183	60	88,24	0,0333
2	mano	0,6730	54	79,41	0,0633
3	nariz	0,6076	54	79,41	0,0932
4	pie	0,5466	53	77,94	0,1226
5	cabeza	0,5306	40	58,82	0,1448
6	brazo	0,5288	47	69,12	0,1709
7	dedo	0,5235	52	76,47	0,1998
8	pierna	0,5073	49	72,06	0,2270
9	uña	0,4826	51	75,00	0,2553
10	oreja	0,4691	44	64,71	0,2797
11	boca	0,4384	39	57,35	0,3013
12	pelo	0,3735	36	52,94	0,3213
13	cuello	0,3652	38	55,88	0,3424
14	rodilla	0,3498	37	54,41	0,3629
15	hombro	0,3032	30	44,12	0,3796
16	diente	0,2865	32	47,06	0,3973
17	corazón	0,2705	32	47,06	0,4151
18	lengua	0,2654	30	44,12	0,4317
19	pecho	0,2595	28	41,18	0,4473
20	cerebro	0,2479	25	36,76	0,4612
21	labio	0,2366	21	30,88	0,4728
22	codo	0,2336	26	38,24	0,4872
23	cara	0,2288	17	25,00	0,4967
24	estómago	0,2206	27	39,71	0,5117
25	pestaña	0,2021	21	30,88	0,5233
26	ceja	0,2001	21	30,88	0,5350
27	espalda	0,1935	25	36,76	0,5488
28	pulmón	0,1881	25	36,76	0,5627
29	hueso	0,1700	20	29,41	0,5738
30	tobillo	0,1582	20	29,41	0,5849
31	cabello	0,1492	14	20,59	0,5927
32	hígado	0,1390	17	25,00	0,6021
33	torso	0,1360	12	17,65	0,6088
34	ombligo	0,1354	14	20,59	0,6165
35	riñón	0,1345	18	26,47	0,6265
36	muslo	0,1345	15	22,06	0,6349

37	piel	0,1317	14	20,59	0,6426
38	muñeca	0,1245	16	23,53	0,6515
39	laringe	0,1199	15	22,06	0,6598
40	mejilla	0,1161	12	17,65	0,6665
41	cadera	0,1135	12	17,65	0,6731
42	intestino	0,1124	16	23,53	0,6820
43	faringe	0,1088	13	19,12	0,6892
44	músculo	0,1041	11	16,18	0,6953

Tabla 60. Vocablos más disponibles del CI *Habilidades docentes* de LH2

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	ojo	0,6763	79	89,77	0,0331
2	mano	0,6154	69	78,41	0,0620
3	cabeza	0,5973	60	68,18	0,0871
4	pie	0,5737	75	85,23	0,1186
5	nariz	0,4961	67	76,14	0,1466
6	brazo	0,4883	59	67,05	0,1713
7	pierna	0,4849	66	75,00	0,1990
8	dedo	0,4571	70	79,55	0,2283
9	oreja	0,4429	62	70,45	0,2543
10	boca	0,4116	55	62,50	0,2773
11	uña	0,3440	59	67,05	0,3021
12	rodilla	0,3104	54	61,36	0,3247
13	pelo	0,2888	42	47,73	0,3423
14	cuello	0,2780	46	52,27	0,3615
15	codo	0,2725	44	50,00	0,3800
16	hombro	0,2701	40	45,45	0,3967
17	diente	0,2291	42	47,73	0,4143
18	lengua	0,1960	36	40,91	0,4294
19	corazón	0,1860	36	40,91	0,4445
20	cerebro	0,1820	33	37,50	0,4583
21	espalda	0,1787	36	40,91	0,4734
22	estómago	0,1687	34	38,64	0,4876
23	pecho	0,1638	32	36,36	0,5010
24	muslo	0,1532	29	32,95	0,5132
25	cadera	0,1517	28	31,82	0,5249
26	tobillo	0,1476	29	32,95	0,5371
27	pulmón	0,1452	30	34,09	0,5496
28	ceja	0,1350	22	25,00	0,5589
29	labio	0,1332	21	23,86	0,5677
30	cara	0,1299	13	14,77	0,5731
31	hígado	0,1277	26	29,55	0,5840
32	hueso	0,1172	27	30,68	0,5953

33	antebrazo	0,1134	19	21,59	0,6033
34	pestaña	0,1090	19	21,59	0,6112
35	talón	0,1055	23	26,14	0,6209
36	muñeca	0,1044	23	26,14	0,6305

3.6.1.8. *Comidas y bebidas*

Comidas y bebidas es el octavo y último centro de interés analizado, como se ha indicado antes, este forma parte del conjunto propuesto por Gougenheim *et al.* (1964) y asumido por el PPHLD. Los resultados de esta investigación muestran que este eje temático es el segundo con mayor cantidad de vocablos con IDL≥0,1 en las tres muestras exploradas. En el caso de los datos aportados por los alumnos de EB, se observan veintisiete unidades léxicas con valores por encima del requerido en la metodología planteada (Tabla 61). En cambio, los datos de LH1 exponen una numeración de treinta y seis lemas (Tabla 62), mientras que los listados digitales de esta misma carrera universitaria suman cuarenta y un vocablos (Tabla 63).

Respecto a las piezas léxicas únicas que encabezan los listados de los tres corpus, se observa que la palabra *Coca-Cola*, que remite a una marca comercial, es la más disponible, tanto en la muestra de Educación Básica como de Letras Hispánicas, alcanzando IDL iguales a 0,4829 (63,89 %) y 0,4645 (55,88 %), respectivamente. En cambio, los resultados de la muestra digital de Letras indican que el vocablo más disponible es *arroz*, que refleja un IDL = 0,4929 y un porcentaje de aparición de 67,05 %.

Tabla 61. Vocablos más disponibles del CI *Comidas y bebidas* de Ed. Básica

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	Coca-Cola	0,4829	69	63,89	0,02870
2	arroz	0,3331	68	62,96	0,05699
3	fideo	0,3010	54	50,00	0,07945
4	papa frita	0,2970	46	42,59	0,09859
5	pizza	0,2917	49	45,37	0,11897
6	jugo	0,2817	51	47,22	0,14018
7	agua	0,2514	46	42,59	0,15932
8	hamburguesa	0,2476	45	41,67	0,17804
9	carne	0,2202	44	40,74	0,19634
10	pollo	0,2049	41	37,96	0,21339
11	sushi	0,1954	35	32,41	0,22795
12	café	0,1720	35	32,41	0,24251
13	lechuga	0,1644	37	34,26	0,25790
14	completo	0,1554	25	23,15	0,26830
15	Pepsi	0,1477	25	23,15	0,27870
16	Sprite	0,1446	26	24,07	0,28952
17	poroto	0,1427	32	29,63	0,30283

18	Fanta	0,1421	24	22,22	0,31281
19	pan	0,1317	28	25,93	0,32446
20	té	0,1255	30	27,78	0,33694
21	cazuela	0,1217	26	24,07	0,34775
22	tomate	0,1182	28	25,93	0,35940
23	verdura	0,1175	24	22,22	0,36938
24	galleta	0,1063	22	20,37	0,37854
25	lenteja	0,1060	27	25,00	0,38977
26	manzana	0,1044	20	18,52	0,39809
27	palta	0,1002	20	18,52	0,40641

Tabla 62. Vocablos más disponibles del CI *Comidas y bebidas* de LH1

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	Coca-Cola	0,4645	38	55,88	0,02141
2	agua	0,4225	44	64,71	0,04620
3	arroz	0,3754	39	57,35	0,06817
4	jugo	0,3687	39	57,35	0,09014
5	carne	0,3015	38	55,88	0,11155
6	fideo	0,2900	31	45,59	0,12901
7	pizza	0,2755	32	47,06	0,14704
8	hamburguesa	0,2631	29	42,65	0,16338
9	té	0,2491	28	41,18	0,17915
10	café	0,2309	28	41,18	0,19493
11	pollo	0,2060	23	33,82	0,20789
12	poroto	0,2036	26	38,24	0,22254
13	lasaña	0,2017	21	30,88	0,23437
14	papa frita	0,2014	18	26,47	0,24451
15	puré	0,1927	21	30,88	0,25634
16	lenteja	0,1865	24	35,29	0,26986
17	cazuela	0,1740	20	29,41	0,28113
18	tomate	0,1681	25	36,76	0,29521
19	papa	0,1580	19	27,94	0,30592
20	Pepsi	0,1571	15	22,06	0,31437
21	queso	0,1571	20	29,41	0,32563
22	pan	0,1570	20	29,41	0,33690
23	lechuga	0,1565	23	33,82	0,34986
24	ensalada	0,1365	17	25,00	0,35944
25	manzana	0,1362	19	27,94	0,37014
26	fruta	0,1319	18	26,47	0,38028
27	tallarín	0,1302	11	16,18	0,38648
28	sushi	0,1293	13	19,12	0,39380
29	cerveza	0,1272	16	23,53	0,40282
30	huevo	0,1241	19	27,94	0,41352

31	leche	0,1228	17	25,00	0,42310
32	Fanta	0,1213	11	16,18	0,42930
33	Sprite	0,1127	11	16,18	0,43549
34	pasta	0,1109	12	17,65	0,44225
35	pescado	0,1060	16	23,53	0,45127
36	garbanzo	0,1052	15	22,06	0,45972

Tabla 63. Vocablos más disponibles del CI *Comidas y bebidas* de LH2

	Vocablo	IDL≥0,1	Frecuencia	% Aparición	Frec. Acumulada
1	arroz	0,4929	59	67,05	0,02489
2	agua	0,4373	54	61,36	0,04768
3	jugo	0,3920	51	57,95	0,06920
4	Coca-Cola	0,3877	45	51,14	0,08819
5	fideo	0,3073	42	47,73	0,10591
6	pan	0,2933	45	51,14	0,12489
7	carne	0,2716	37	42,05	0,14051
8	té	0,2567	36	40,91	0,15570
9	papa	0,2515	34	38,64	0,17004
10	lechuga	0,2367	29	32,95	0,18228
11	poroto	0,2342	32	36,36	0,19578
12	café	0,2241	33	37,50	0,20970
13	fruta	0,2145	28	31,82	0,22152
14	tomate	0,1953	33	37,50	0,23544
15	pollo	0,1918	29	32,95	0,24768
16	hamburguesa	0,1891	27	30,68	0,25907
17	verdura	0,1808	24	27,27	0,26920
18	lenteja	0,1771	25	28,41	0,27975
19	pizza	0,1753	24	27,27	0,28987
20	Sprite	0,1632	22	25,00	0,29916
21	leche	0,1604	27	30,68	0,31055
22	tallarín	0,1578	20	22,73	0,31899
23	cazuela	0,1567	24	27,27	0,32911
24	puré	0,1561	21	23,86	0,33797
25	Fanta	0,1559	20	22,73	0,34641
26	huevo	0,1529	26	29,55	0,35738
27	ensalada	0,1415	20	22,73	0,36582
28	papa frita	0,1341	18	20,45	0,37342
29	helado	0,1299	19	21,59	0,38143
30	palta	0,1261	20	22,73	0,38987
31	garbanzo	0,1254	18	20,45	0,39747
32	queso	0,1241	24	27,27	0,40759
33	galleta	0,1221	19	21,59	0,41561
34	completo	0,1206	18	20,45	0,42321
35	bebida	0,1185	14	15,91	0,42911

36	pescado	0,1172	18	20,45	0,43671
37	Pepsi	0,1166	14	15,91	0,44262
38	sushi	0,1142	18	20,45	0,45021
39	salsa	0,1034	17	19,32	0,45738
40	cerveza	0,1022	15	17,05	0,46371
41	frutilla	0,1016	17	19,32	0,47089

En el siguiente capítulo, se describe y explican los resultados de los análisis correlacionales entre el léxico disponible y las variables socioeducativas propuestas.

Capítulo 4. Análisis sociolingüístico

4.1. Consideraciones previas

El término *investigación científica* puede activar en la mente palabras como: *sujetos*, *datos*, *corpus*, *matemática*, *fórmulas*, *cálculo*, *computadora*, *programas*, *gráficos*, *deducción*, *inducción*, etc., hasta algunas voces nada o poco relacionadas directamente con el concepto evocador, como: *difícil*, *arduo*, *miedo*, *fuerza*, *tiempo*, *dedicación*, etc., creando una red léxica asociativa en torno a él. Este ejemplo hipotético ha intentado reflejar una pequeña arista de lo que significa el trabajo intelectual vinculado con las prácticas investigativas. A su vez, abre paso a este capítulo en el que se exponen los análisis cuantitativos, de corte sociolingüístico, del léxico disponible de los estudiantes de Educación Básica y Letras Hispánicas, UC.

Entonces, la investigación científica puede definirse como "un proceso de indagación sistemática, cuya característica principal es la replicabilidad, en el que se recogen, analizan, interpretan y utilizan los datos con la finalidad de comprender, describir, predecir y explicar el fenómeno que nos planteamos" (Herrera *et al.* 2016: 13). Es decir, a partir de un interés genuino de un especialista hacia un fenómeno particular de cualquier naturaleza –en este caso, lingüístico–, se observan esquemática y analíticamente los datos empíricos que se tienen del objeto bajo estudio, tratándolo y procesándolo mediante técnicas y herramientas convencionalmente aceptadas por una comunidad discursiva, con la intensión de conocer y, por supuesto, informar y exponer de manera clara las cualidades descubiertas del fenómenos en cuestión, ofreciendo la posibilidad de que pueda replicarse en otras poblaciones o conjunto de datos similares.

A partir de la panorámica brindada en el acápite previo, puede acotarse que un grueso de los estudios científicos empiristas, como los desarrollados desde algunas vertientes de la fonética, psicolingüística, lexicometría, entre otras disciplinas y ciencias, se aplican métodos y técnicas provenientes de la estadística. Esta se entiende, a grandes rasgos, como una herramienta que permite tomar decisiones y argumentar respecto a un hecho, con base en la información cuantitativa o matemática que se tenga a partir de los datos explorados. En palabras de Seoane *et al.* (2007: 466), la estadística "[…] se define como la ciencia matemática que se refiere a la recopilación, estudio e interpretación de los datos obtenidos en un estudio". Esta conceptualización parece ajustarse, entonces, a las propuestas metodológicas de algunos trabajos sociolingüísticos de corte cuantitativo, en los que, a partir de un corpus representativo de una comunidad de habla, se observan los datos de un fenómeno lingüístico dado.

Hogaño, la estadística cuenta con la computación como una gran aliada, ya que, como afirma Mafokosi (2009: 17) gracias a esta última "se abre el horizonte de las técnicas analíticas avanzadas a cualquiera que quiere explorar nuevas preguntas o pasar revistas a las antiguas". De manera que la creación de programas y aplicaciones computarizados no solo ha permitido desarrollar trabajos cuantitativos de forma más rápida y confiable, sino también ha ampliado el acceso al mundo de las estadísticas.

En este contexto, la disponibilidad léxica es una de las líneas de investigación dentro del área de la lingüística que recurre a la estadística para analizar e interpretar los fenómenos léxicos de una sintopía o, en este caso, una comunidad discursiva particular. En este tipo de estudios léxico-métricos, generalmente, se realizan dos tipos de análisis cuantitativos, a saber: descriptivo e inferencial.

La estadística descriptiva procura exponer, de manera ordenada y sintética, las cualidades numéricas de los datos recogidos de un conjunto mayor o población. Esta es, a tenor de lo afirmado por Seoane *et al.* (2007: 467), "[…] la parte de la estadística que sintetiza y resume la información contenida en un conjunto de datos, por tanto, un análisis descriptivo consiste en clasificar, representar y resumir los datos". Por su parte, la estadística inferencial busca encontrar y determinar predicciones o inferencias desde los datos observados, para lo que se fundamenta en análisis construidos sobre la base de la teoría probabilística. De forma más elaborada, Seoane *et al.* (2007: 469) indican que:

> La inferencia se define como el conjunto de métodos estadísticos que permiten deducir cómo se distribuye la población e inferir las relaciones entre variables a partir de la información que proporciona la muestra recogida. Por tanto, los objetivos fundamentales de la inferencia estadística son la estimación y el contraste de hipótesis.

En consonancia con lo anterior, puede señalarse que los resultados inferenciales, basados en una muestra pequeña, pueden proyectarse a la población general de donde emergieron los casos analizados. Sin embargo, antes de realizar los cálculos estadísticos inferenciales, los datos deben pasar primero por los análisis estadísticos descriptivos (Mafokosi, 2009).

En virtud de uno de los objetivos de esta tesis doctoral que reza determinar la relación entre el léxico disponible de los informantes y las categorías sociológicas y socioeducativas, y, en consonancia con los antecedentes, se ha procedido a efectuar un conjunto de análisis cuantitativos escalonadamente, desde los cálculos univariantes hasta los bivariantes. Para lograr dichas metas, se han tomado en cuenta las cualidades de cada una de las variables propuestas en esta pesquisa. La variable dependiente –entiéndase, aquella que sufre algún cambio cuando es expuesta o entra en interacción con las propiedades de otras variables– corresponde al caudal léxico de los encuestados, respecto a los ocho centros de interés planteados. Por su parte, las variables independientes –a saber,

los factores capaces de incidir en mayor o menor grado sobre otras variables– aplicadas en esta tesis son: *Sexo, Carrera, Año o nivel de curso, Cantidad de libros leídos, Frecuencia semanal de lectura optativa* y *Tipo de formato.*

Se ha empezado con la descripción individual de cada una de las variables a partir de los análisis descriptivos, que apuntaban a determinar la distribución de cada factor. Como se ha indicado en la metodología, estos seis factores se caracterizan por ser dicotómicos o politómicos. Por esta razón, se han aplicado test paramétricos distintos, basados en la comparación de medias, que es "una de las técnicas más frecuentes y útiles en la investigación sociológica" (Gómez Devís, 2004: 128).

En los casos de las variables de dos variantes o niveles (dicotómicas) se ha optado por la prueba paramétrica t de Student, creada por William Sealy Gosset, y se refiere a "un conjunto de curvas estructurada por un grupo de datos de unas muestras en particular" (Sánchez, 2015: 59). Es requisito indispensable que las dos muestras independientes a las que se les quiere determinar la significación de las diferencias observadas "tengan distribución normal y homogeneidad en sus varianzas" (Sánchez, 2015: 59). Por su parte, para los análisis de los factores politómicos (de 3 y más variantes) se apeló a la prueba de Anova de un factor. A la que Blanca *et al.* (2017: 552) definen como: "The F-test assumes that the outcome variable is normally and independently distributed with equal variances among groups". Ambas pruebas permiten determinar la significación estadística de la relación entre una variable dependiente y otra independiente. En lingüística y, por ende, en lexicometría se ha convenido asumir que la relación de los datos explorados es significativa cuando el estadístico utilizado arroja un nivel de significación estadística o valor de $p = 0{,}05$.

Con el fin de no excluir de los análisis inferenciales los grupos cuyas medias no cumplieran el requisito de homocedasticidad o igualdad de varianza, se apeló a los test no paramétricos adecuados para el tipo de muestra. En atención a lo cual, para las categorías dicotómicas se empleó la prueba U de Mann-Whitney, puesto que permite verificar "la H_0 de que 2 muestras aleatorias autónomas provienen de dos poblaciones iguales o de una misma población, cuando no se cumple el supuesto de normalidad y homocedasticidad, medidos mínimo en escala ordinal" (Ramírez y Polack, 2020: 197). En cambio, para las categorías politómicas se administró la H de Kruskal-Wallis, que "sirve para probar la H_0 de que las m muestras autónomas provienen de poblaciones similares o de una misma población, aquí la variable que se estudia tiene una distribución continua, con escala mínimamente ordinal" (Ramírez y Polack, 2020: 201).

Como se ha indicado en el apartado 2.4 de la metodología, los participantes del estudio son en total 264 estudiantes de las carreras de Educación Básica (EB) y Letras Hispánicas (LH) de la PUC,

cuyas listas de palabras han conformado el corpus general. A partir de dicho total, se organizaron tres muestras diferentes, que han sido definidas a partir de los factores *Formato de prueba* y *Carrera*.

A manera de recapitulación, los dos primeros conjuntos de datos están conformados por los listados de palabras recogidos mediante el método tradicional o formato en papel, la primera muestra concierne a los datos suministrados por estudiantes de la carrera de Educación Básica. Por su parte, la segunda muestra corresponde a las respuestas de los alumnos de Letras Hispánicas. Por último, la tercera muestra también atañe a discentes humanistas, pero se distingue en que los datos se recolectaron a través de la página web creada para esta tesis, es decir, se refiere al método alternativo propuesto. En conformidad con lo anterior, en los siguientes apartados se exponen los análisis cuantitativos del léxico disponible de cada una de las tres muestras.

4.2. Estadística descriptiva de las variables sociológicas

4.2.1. Sexo

El sexo es una de las variables sociodemográficas más recurrentes en los estudios de disponibilidad léxica, sobre todo en los que se desarrollan desde una perspectiva sociolingüística. Este es uno de los factores sociales utilizados para analizar el caudal léxico de los 264 estudiantes universitarios del corpus. En líneas generales, los participantes son mujeres en su mayoría, 208 entrevistadas, lo que representa el 78,8 % de los datos; es decir más de las ¾ partes del corpus. Por su parte, los 56 restantes son hombres, el 21,2 %, como se aprecia en la tabla 64.

Tabla 64. Distribución de los 264 participantes del estudio, según *Sexo*.

		Frecuencia	Porcentaje
Válidos	Hombres	56	21,2
	Mujeres	208	78,8
	Total	264	100,0

Las respuestas de estos 264 participantes se encuentran organizadas en dos subcorpus, definidos a tenor de las variables *Carrera* y *Formato de pruebas*, como se ha descrito en el subapartado 2.4. El subcorpus 1 está integrado por los datos de 176 sujetos, mientras que el subcorpus 2 está conformado por las respuestas de 156 informantes. La distribución de estos grupos, conforme con la variable *Sexo*, se detalla en las tablas 65 y 66.

Tabla 65. Distribución, según la variable *Sexo*, de los informantes del subcorpus 1.

		Frecuencia	Porcentaje
Válidos	Hombres	38	21,6
	Mujeres	138	78,4
	Total	176	100,0

Tabla 66. Distribución, según la variable *Sexo*, de los informantes del subcorpus 2.

		Frecuencia	Porcentaje
Válidos	Hombres	38	24,4
	Mujeres	118	75,6
	Total	156	100,0

Como puede observarse, en ambos subcorpus los informantes son mayormente mujeres, cuyos porcentajes superan las ¾ de los datos. En el subcorpus 1 el número de féminas es 138, lo que representa el 78,4 %. En tanto, los hombres son 38, un 21,6 % de los informantes. Por su parte, el número de mujeres en el subcorpus 2 es 118, lo que se traduce en un 75,6 % de los entrevistado, mientras que los hombres son 38, el 24,4 % de los encuestados.

En virtud de que el subcorpus 1 está integrado por datos de estudiantes de dos Facultades, a continuación, se desglosan sus características según el sexo. De los 108 individuos de la carrera de Educación Básica, se contabilizan 90 mujeres, un 83,3 %, y 18 hombres, lo que representa un 16,7 % de los participantes. A la par, de los 68 sujetos de Letras Hispánica, 48 son mujeres, alcanzando un 70,6 %, y los hombres suman 20 encuestados, un 29,4 %. Estos datos se leen en las Tablas 67 y 68.

Tabla 67. Distribución, según la variable *Sexo*, de los informantes de Ed. Básica

		Frecuencia	Porcentaje
Válidos	Hombres	18	16,7
	Mujeres	90	83,3
	Total	108	100,0

Tabla 68. Distribución, según la variable *Sexo*, de los informantes de LH1

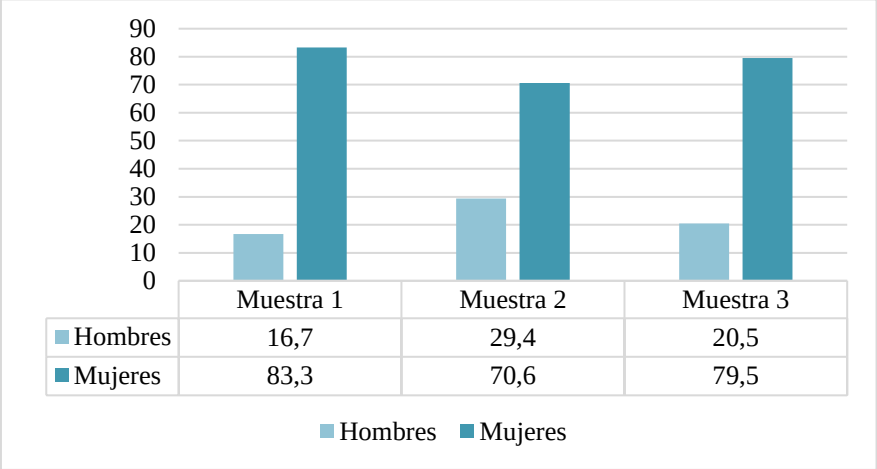
		Frecuencia	Porcentaje
Válidos	Hombres	20	29,4
	Mujeres	48	70,6
	Total	68	100,0

En el caso del subcorpus 2, únicamente queda por detallar las cualidades de los sujetos que realizaron los test de manera *online*. Estos suman 88 individuos en total, de los cuales 70 son mujeres, que corresponde al 79,5 % de la muestra, y 18 son hombres, lo que representa el 20,5 %. Estos números se aprecian en la Tabla 69.

Tabla 69. Distribución, según la variable Sexo, de los informantes de LH1

		Frecuencia	Porcentaje
Válidos	Hombre	18	20,5
	Mujer	70	79,5
	Total	88	100,0

Gráfico 13. Comparación del número de informantes por Sexo en las tres muestras.



Como se observa en el Gráfico 13, a pesar de las disparidad numérica y porcentual, referente al sexo de los participantes en cada muestra, se observa un mismo patrón distribucional, en el que las mujeres superan a los hombres, por encima de la barra del 60 %. Contrariamente, los hombres no alcanzan si quiera la línea cuadrícula del 30 %.

4.2.2. Carrera

El objetivo de esta variable es contrastar las convergencias y divergencias –cuantitativas y cualitativas– del léxico disponible de estudiantes de Educación Básica y Letras Hispánicas de la UC. El primero programa se encuentra inscrito en la Facultad de Educación y el segundo, en la Facultad de Letras. Se considera que ambas carreras parecen compartir algunas parcelas de conocimiento, especialmente en los temas relacionados con la lengua y la lectura. Sin embargo, cada una los estudias desde los enfoques relativos a sus respectivas áreas del saber. En este sentido, analizar el léxico disponible de los estudiantes de ambas carreras supondría que habrá un alto componente léxico compartido, pero también cabría observar un vocabulario diferenciador. En consonancia con lo anterior, esta variable está integrada por dos variantes, que corresponden cada una a las carreras observadas. En las Tabla 70, se expone la distribución de los encuestados en función de esta variable.

Tabla 70. Distribución, según la variable *Carrera*, de los participantes del subcorpus 1.

		Frecuencia	Porcentaje
Válidos	Ed. Básica	108	61,4
	Letras	68	38,6
	Total	176	100,0

Puede observarse que más de la mitad de los encuestados del subcorpus 1 estudiaban Educación Básica, concretamente 108, lo que representa el 61,4 % de la muestra. En cuanto a los entrevistados que cursaban Letras Hispánicas, estos son 68 individuos, lo que se traduce en un 38,6 % del corpus. Entre ambos grupos hay 22,8 puntos porcentuales de separación. Es decir, existe una gran diferencia en la distribución de los datos por sujetos en virtud de las áreas de especialidad.

4.2.3. Año o nivel de curso

La variable *Año o nivel de estudio* ha resultado estadísticamente significativa en algunos trabajos de disponibilidad léxica (cf. Herranz, 2018, 2020), y ha mostrado que los estudiantes más avanzados reportan más unidades léxicas que los de los grados menores del sistema educativo. En esta tesis se ha utilizado este factor con los objetivos de, en primer lugar, determinar la incidencia del nivel universitario de los informantes en el caudal léxico. En segundo lugar, comparar los lexicones de estudiantes de recién ingreso a la universidad con los de quienes están en niveles avanzados. En este sentido se supondría que: i) los sujetos con más año de escolaridad universitaria tendrían un mayor caudal léxico que los individuos que están iniciando sus carreras; ii) el léxico disponible de los alumnos de los últimos años de universidad presentaría una mayor grado de cohesión que el vocabulario de quienes cursan el primer año de carrera; y iii) si bien pudiera haber mayor número de lexías de conocimiento general, los discentes más avanzados evocarían términos más técnicos o disciplinares. En este orden de ideas, en esta tesis la variable *Año de curso* está constituida por las siguientes dos variantes: i) 1.er Año y ii) 4.º Año.

El análisis descriptivo arroja que, en el subcorpus 1 (Tabla, de los 176 informantes, 117 son de 1.er Año, lo que representa el 66,5 % de la muestra, mientras que 59 son de 4.º Año, el 33,5 %. Al desglosar los datos de este subcorpus por las carreras de los participantes, el análisis indicó que, de los 108 cursantes de Educación Básica, 83 se encontraban en 1.er Año (76,9 %) y los restantes 25, en el 4.º Año (23,1 %). En el caso del grupo de Letras Hispánicas, la distribución es equitativa, 34 en cada nivel. En cuanto al grupo que realizó las pruebas *online*, a saber: 66 estudiantes, 39 de ellos cursaban 1.er Año (44,3 %), y 49, el 4. ° Año (55,7 %). En las Tablas 71, 72 y 73 pueden verse estos resultados.

Tabla 71. Participantes por Año de curso del subcorpus 1.

		Frecuencia	Porcentaje
Válidos	1.er Año	117	66,5
	4.º Año	59	33,5
	Total	176	100,0

Tabla 72. Participantes por *Carrera* y *Año de curso* del subcorpus 1.

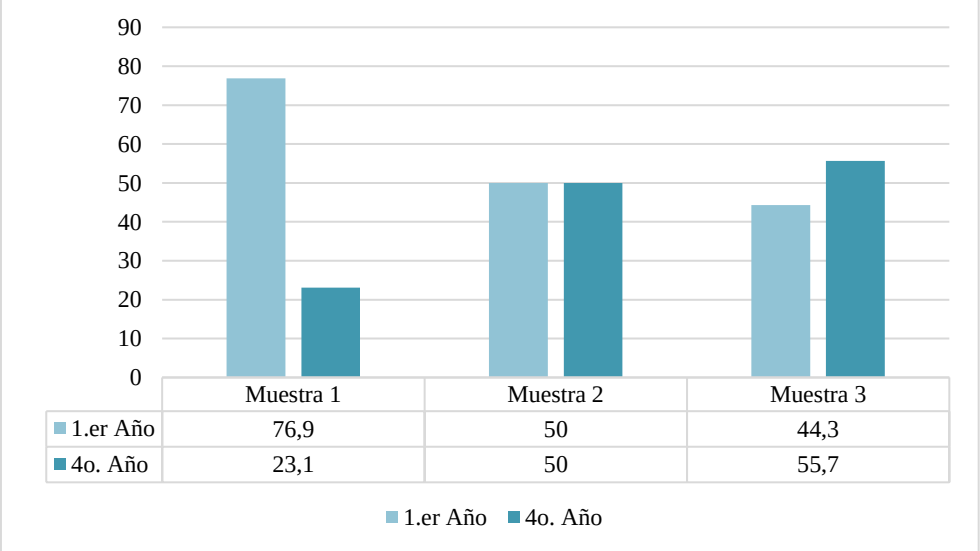
		Años de curso				Total
		1.er Año	%	4.º Año	%	
Carrera	Ed. Básica	83	70,9	25	42,4	108
	Letras	34	29,1	34	57,6	68
	Total	117		59		176

Tabla 73. Participantes por Año de curso de la muestra digital

		Frecuencia	Porcentaje
Válidos	1.er Año	39	44,3
	4.º Año	49	55,7
	Total	88	100,0

Al contrastar los cómputos de cada muestra, se aprecian patrones completamente disímiles. En efecto, en el Gráfico 14 se aprecia que la distribución desigual de los participantes, según el nivel de curso, de la muestra 1 (Educación Básica) se inclina desproporcionalmente hacia la variante 1.er Año, mostrando que la barra supera la línea de la cuadrícula de 70 %. Contrariamente, la muestra digital, en la que los casos por año de curso también son dispares, la barra más alta corresponde a los datos de la variante 4.º Año. Por último, la muestra 2 (Letras Hispánica, formato físico) es la única que tiene una ordenación equitativa de los datos correspondientes al nivel de curso de los participantes.

Gráfico 14. Comparación del número de informantes por formato de prueba y año de curso



4.2.4. Formato de pruebas

Uno de los objetivos principales de esta investigación es evaluar la aplicación de los test de disponibilidad léxica a través de un método alternativo al que usualmente se emplea en este tipo de trabajos. La técnica a la que sea recurrido en esta tesis es digital. Es decir, se construyó una página web que permitiera recopilar léxico disponible de la manera más parecida posible a la metodología planteada en el PPHLD, como se ha explicado en el epígrafe 2.6. Se partía del supuesto de que los datos recopilados digitalmente no distarían tanto, cuantitativa ni cualitativamente, de los tomados en papel. No obstante, por razones externas a la investigación, los test computarizados fueron respondidos solamente por alumnos de Letras Hispánicas. En virtud de lo cual, los análisis contrastivos del LD recogido de forma remota se llevaron a cabo con los vocabularios de los discentes de la Facultad de Letras. De esta manera se mantuvo el criterio de homogeneidad de la población. En la tabla 74, se exhibe la distribución de los encuestados según la variable *Formato de pruebas*.

Tabla 74. Encuestados según *Formato de pruebas*.

		Frecuencia	Porcentaje
Válidos	Tradicional	68	43,6
	Digital	88	56,4
	Total	156	100,0

El subcorpus 2 consta de las listas de palabras, recogidas en papel y digital, de 156 estudiantes de Letras. De estos, 88 respondieron los cuestionarios mediante la plataforma web, lo que representa el 56,4 % de la muestra; mientras que los demás 68 discentes contestaron los test de forma tradicional: en papel y presencial, esto comprende el 43,6 % de los participantes. Entre ambos grupos hay una diferencia porcentual de 12,8 puntos.

4.2.5. Cantidad de libros leídos en el último año

Cantidad de libros leídos[17]es una variable referida a consumos culturales, ha sido poco utilizada en los trabajos de DL, según el arqueo bibliográfico. Algunos de los investigadores que la han aplicado son Ávila-Muñoz (2007) y Santos Díaz (2020). El objetivo de este factor es conocer la incidencia que tiene el número de libros leídos en el léxico disponible de los encuestados. Se parte del supuesto de que los estudiantes que leyeron más libros en un año tendrían un mayor caudal léxico que aquellos quienes no leyeron nada o lo hicieron poco. Esta variable consta de cuatro variantes o niveles, a saber: 1) Ningún libro, 2) De 1 a 5 libros, 3) De 6 a 10 libros, y 4) Más de 10 libros.

[17] En el texto se usará el nombre apocopado de esta variables: *Cantidad de libros leídos*.

En el subcorpus 1, los datos indican que, de 176 sujetos, 76 seleccionaron la variante 2 (De 1 a 5 libros), lo que representa el 43,2 %; 45 informantes escogieron la variante 4 (Más de 10 libros), un 25,6 %; 38 (21,6 %) se decantaron por la variante 3 (De 6 a 10 libros) y 17, el 9,7 % de los participantes, optaron por la variante 1 (Ninguno). En la tabla 75 se muestran estos cómputos.

Tabla 75. Distribución de los datos del subcorpus 1, según Cantidad de libros leídos.

		Frecuencia	Porcentaje
Válidos	Ninguno	17	9,7
	De 1 a 5	76	43,2
	De 6 a 10	38	21,6
	Más de 10	45	25,6
	Total	176	100,0

Sobre los resultados de los alumnos de Educación Básica, en la Tabla 76 se aprecia que, de los 108 participantes, 9 indicaron no haber leído ningún libro (8,3 %); 56 señalaron que leyeron entre 1 y 5 libros (51,9 %); 44 afirmaron haber leído entre 6 y 10 libros (25,9 %); y 15 marcaron la opción Más de 10 libros (13,9 %). Por su parte, los 68 participantes de LH1 (Tabla 77) se distribuyen de la siguiente manera: 8 (11,8 %) indicaron no haber leído ningún libro; 30 afirmaron haber leído más de 10 (44,1 %); 20 optaron por la variante 2 (29,4 %); y 10, por la variante 3 (14,7 %). En la Tabla 78 se leen estos últimos datos. Por último, los datos de la muestra digital arrojan que solo 5 sujetos (5,7 %) afirmaron que no leyeron ningún libro; 42 participantes señalaron que leyeron de 1 a 5 libros (47,7 %); 25 indicaron que leyeron de 6 a 10 libros (28,4 %); y 16 seleccionaron la opción de más de 10 libros (18,2 %). En la Tabla 78 se ilustran estos datos.

Tabla 76. Distribución de los datos de Ed. Básica, según Cantidad de libros leídos.

		Frecuencia	Porcentaje
Válidos	Ninguno	9	8,3
	De 1 a 5	56	51,9
	De 6 a 10	28	25,9
	Más de 10	15	13,9
	Total	108	100,0

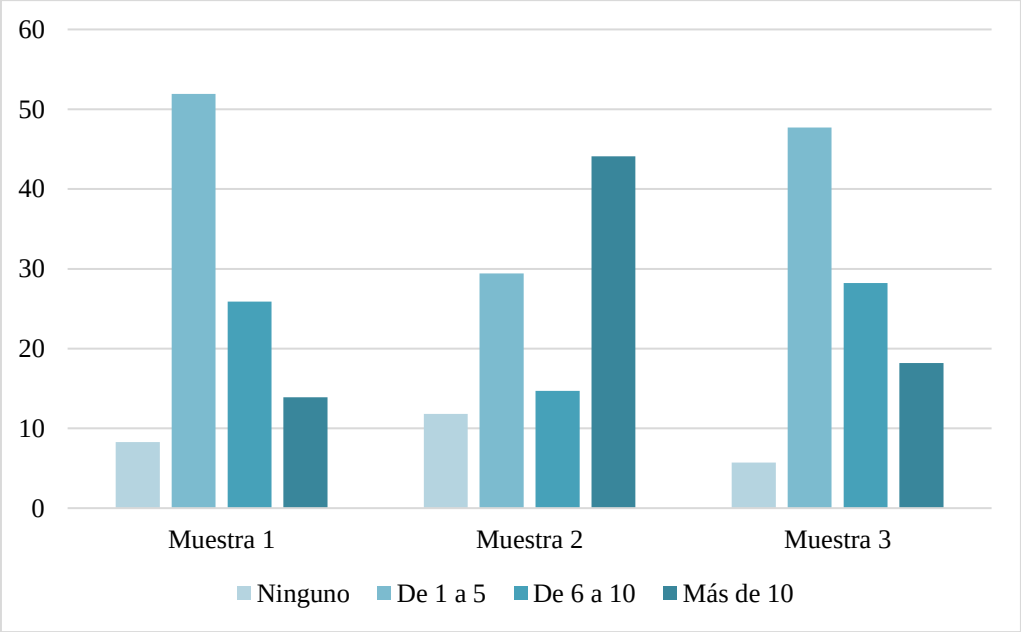
Tabla 77. Distribución de los datos de LH1, según Cantidad de libros leídos.

		Frecuencia	Porcentaje
Válidos	Ninguno	8	11,8
	De 1 a 5	20	29,4
	De 6 a 10	10	14,7
	Más de 10	30	44,1
	Total	68	100,0

Tabla 78. Distribución de los datos De LH2, según Cantidad de libros leídos.

		Frecuencia	Porcentaje
Válidos	Ninguno	5	5,7
	De 1 a 5	42	47,7
	De 6 a 10	25	28,4
	Más de 11	16	18,2
	Total	88	100,0

Gráfico 15. Comparación de la distribución de los informantes según Cantidad de libros leídos.



En el Gráfico 15, llama la atención que las muestras 1 y 3 presenten un patrón similar en la distribución de los datos de los participantes conforme a la variable *Cantidad de libros leídos*, en el que se aprecia que la barra que representa la variante De 1 a 5 libros es la sobresaliente, indicando la tendencia de los entrevistados respecto al consumo cultural analizado. Asimismo, se observa que la mayoría de los datos de ambos grupos se aglutinan en las variantes 2 y 3. En cambio, la muestra 2 expone un modelo completamente dispar al de los otros dos conjuntos; en el que la tendencia parece no estar clara. En este la barra que prevalece es la referida al nivel 4 de la variante (Más de 10 libros), pero la segunda más predominante es la del nivel 2.

4.2.6. Frecuencia semanal de lectura optativa

La última variable analizada es *Frecuencia de lectura*[18], la cual toma en consideración el tiempo semanal que, en promedio, los sujetos dedican a la lectura de textos optativos o a la realización

[18] En el cuerpo del texto, el nombre de esta variable se ha apocopado en *Frecuencia de lectura*.

de lectura por placer. En este caso se puso el foco en el acercamiento que los participantes pudieran tener a una literatura distinta a la exigida por sus respectivas mallas de estudios. La pregunta que se les planteó fue: *En promedio, ¿cuántas horas a la semana dedicas a la lectura de textos optativos?* Las respuestas eran de tipo cerrada, por lo que los encuestados debían escoger una de estas cuatro opciones: 1) Ninguna hora, 2) De 1 a 5 horas, 3) De 6 a 10 horas, y 4) Más de 10 horas a la semana. Esta es una variable de tipo categorial, de manera que las cuatro respuestas corresponden a los niveles de análisis o variantes.

Al analizar el subcorpus 1, compuesto por 176 informantes de Educación y Letras, se observa que más de la mitad de los datos, las respuestas de 119 (67,6 %), se concentra en la variante 2 (De 1 a 5 horas); 23 respuestas se hallan en la variante 1 (13,1 %); las variantes 3 aglutina los datos de 21 sujetos (11,9 %), y la 4, los de 13 (7,4 %), como se detalla en la Tabla 79.

Tabla 79. Distribución de los datos del subcorpus 1, según Frecuencia de lectura.

		Frecuencia	Porcentaje
Válidos	Ninguna	23	13,1
	De 1 a 5	119	67,6
	De 6 a 10	21	11,9
	Más de 10 horas	13	7,4
	Total	176	100,0

Al desglosar los datos del subcorpus 1, los resultados indican que, las respuestas de los 108 alumnos de la Facultad de Educación se reparten de la siguiente manera: 80 (74,1 %) en la segunda variante (De 1 a 5 horas); 15 en la primera (Ninguna hora), un 13,9 %; 9 en la tercera (De 6 a 10 horas), 8,3 %; y 4 en la cuarta, apenas un 3,7 %. Por su parte, los datos de los 68 encuestados de la Facultad de Letras se distribuyen como sigue: 39 en la segunda variante, lo que se traduce en 57,4 %; 12 en la tercera opción (17,6 %); 8 en la cuarta variante (11,8 %) y 9 en la primera (13,2 %). Por último, los datos de la muestra digital se distribuyen de la siguiente manera: de los 88 participantes del estudio, 55 optaron por el segundo nivel de análisis (62,5 %); 15 se decantaron por el primero (17 %); 13 (14,8 %) y 5 (5,7 %) encuestados seleccionaron los niveles 3 y 4, respectivamente. Estos cómputos pueden verse en las tablas 80, 81 y 82.

Tabla 80. Distribución de los estudiantes de EB, según *Frecuencia de lectura*

		Frecuencia	Porcentaje
Válidos	Ninguna	15	13,9
	De 1 a 5 horas	80	74,1
	De 6 a 10 horas	9	8,3
	Más de 10 horas	4	3,7
	Total	108	100,0

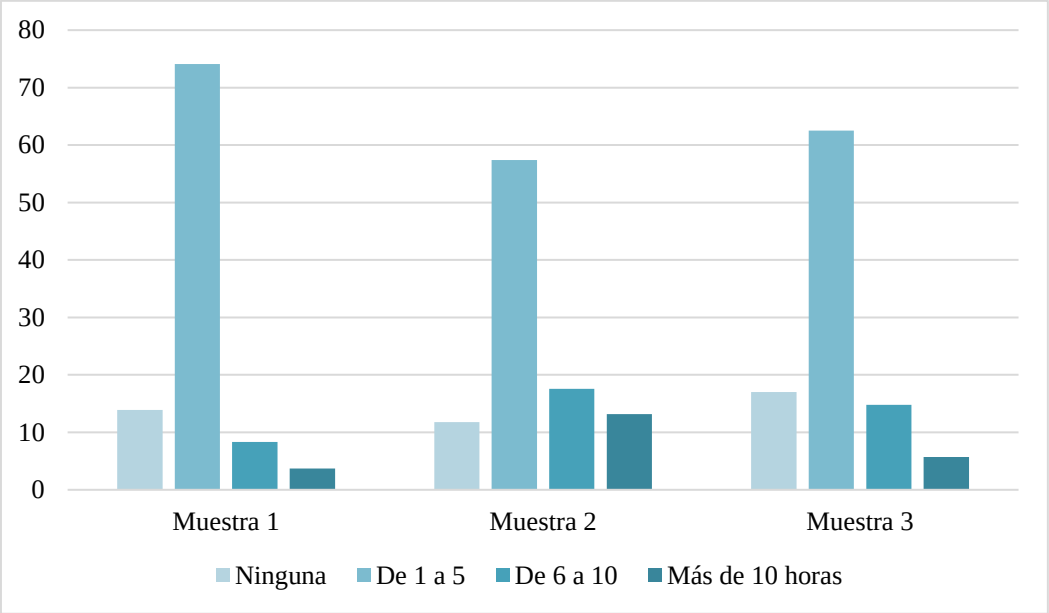
Tabla 81. Distribución de los estudiantes de LH1, según *Frecuencia de lectura*

		Frecuencia	Porcentaje
Válidos	Ninguna	8	11,8
	De 1 a 5 horas	39	57,4
	De 6 a 10 horas	12	17,6
	Más de 10 horas	9	13,2
	Total	68	100,0

Tabla 82. Distribución de los estudiantes de LH2, según *Frecuencia de lectura*

		Frecuencia	Porcentaje
Válidos	Ninguna	15	17,0
	1 a 5 horas	55	62,5
	De 6 a 10 horas	13	14,8
	Más de 10 horas	5	5,7
	Total	88	100,0

Gráfico 16. Comparación de la distribución de los informantes según Frecuencia de lectura.



Las tres muestras analizadas concentran las respuestas acerca de la frecuencia de lectura en el segundo nivel analítico, sobrepasando la línea del 50 %, lo que evidencia que los encuestados tienen una muy baja práctica por la lectura optativa. Contrastivamente, los datos de los dos corpus de Letras presentan patrones casi similares, pero diferenciados por el mayor acopio de casos en la variante 1 (Ninguna hora) apreciable en la muestra digital. En este sentido, el grupo de Educación Básica se ajusta más al modelo de distribución de datos reflejado por el corpus digital.

4.3. Análisis bivariantes

Una vez que se ha descrito cuantitativamente la distribución de los datos en función de las variables extralingüísticas (*Sexo*, *Carrera*, *Año de curso*, *Formato de pruebas*, *Cantidad de libros leídos* y *Frecuencia de lectura*) corresponde, en consonancia con los objetivos de esta tesis doctoral, analizar la relación estadística entre dichos factores y el caudal léxico de los encuestados. Así pues, el objetivo de esta sección es determinar la incidencia de los factores en el léxico disponible de los estudiantes universitarios de Educación Básica y Letras Hispánicas de la UC, de manera que pueda contarse con una descripción sociolingüística de los lexicones explorados. Para esto se ha recurrido a la aplicación de las pruebas paramétricas t de Studen, cuando la variable es dicotómica, y Anova, cuando es politómicas, a través del programa computarizado SPSS.

4.3.1. Sexo

En este epígrafe se examina la relación entre el *Sexo* y el caudal léxico de los encuestados, organizado en tres muestras, por lo cual se parte de la observación de la productividad léxica de los grupos, para luego realizar el análisis inferencial correspondiente. Este factor ha resultado significativo en investigaciones como las de Gómez Devís (2004), Gómez Molina y Gómez Devís (2004), Hernández Muñoz (2006) y Mateo García (1998), pero no en otras como las de Galloso (2002) Pacheco *et al.* (2017) y Jiménez (2019). Se parte del supuesto –que debe corroborarse a partir de los análisis estadísticos– de que el sexo de los encuestados (estudiantes universitarios de Educación Básica y Letras Hispánicas) no incide en el léxico disponible.

En la Tabla 83, concerniente a los datos de los universitarios de Educación Básica, se observa que la media global de palabras de las mujeres ($\overline{X}$ = 16,81) supera a la de los hombres ($\overline{X}$ = 15,04), pero en apenas 1,77 puntos de diferencias. Asimismo, las futuras maestras presentan los promedios más altos en todas las áreas nocionales analizadas, siendo *Partes del cuerpo* ($\overline{X}$ = 22,92), *Comidas y bebidas* ($\overline{X}$ = 22,66) y *La escuela* ($\overline{X}$ = 18,71) las que sobrepasan la media general. Estos actualizadores son los que exponen los cómputos por encima de la media del sociolecto masculino, CI07 ($\overline{X}$ = 21,61), CI08 ($\overline{X}$ = 20,11) y CI05 ($\overline{X}$ = 15,44). Sin embargo, ambos grupos exhiben promedios muy similares en el CI06. *Habilidades docentes* (Hombre = 12,56 y Mujeres = 12,61).

Tabla 83. Comparación de la media de palabras entre hombres y mujeres de EB

Centro de Interés	Hombres	Mujeres
01. La lectura	11,17	13,38
02. El profesor	12,28	14,04
03. La educación	13,89	16,13
04. Juegos y distracciones	13,28	14,00
05. La escuela	15,44	18,71
06. Habilidades docentes	12,56	12,61
07. Partes del cuerpo	21,61	22,92
08. Comidas y bebidas	20,11	22,66
Media aritmética	15,04	**16,81**

Contrariamente, en la muestra de Letras Hispánica recogida en papel, se aprecia que son los hombres los que tienen el mayor promedio de palabras global ($\overline{X}$ = 21,18) *versus* las mujeres ($\overline{X}$ = 19,10), pero esta distinción es de apenas 2,08 unidades léxicas. Igualmente, los chicos también superan a las féminas en los ochos actualizadores, de los cuales *Partes del cuerpo* ($\overline{X}$ = 29,75) y *Comidas y bebidas* ($\overline{X}$ = 27,70) se hallan por encima de la media general del conjunto masculino, mientras que los valores de *La escuela* ($\overline{X}$ = 20,30) y *La lectura* ($\overline{X}$ = 20,15) se encuentran cercanos a la global. Por su parte, las mujeres exponen los promedios más elevados en los centros de interés, de mayor a menor, *Comidas y bebidas* ($\overline{X}$ = 25,46), *Partes del cuerpo* ($\overline{X}$ = 25,15) y *La escuela* ($\overline{X}$ = 20,10). Estos datos se detallan en la tabla 84.

Tabla 84. Comparación de los $\overline{X}$ por sexo de LH1

Centro de Interés	Hombres	Mujeres
01. La lectura	20,15	17,98
02. El profesor	18,70	16,46
03. La educación	19,10	17,15
04. Juegos y distracciones	18,70	17,19
05. La escuela	20,30	20,10
06. Habilidades docentes	15,00	13,31
07. Partes del cuerpo	29,75	25,15
08. Comidas y bebidas	27,70	25,46
Media aritmética	**21,18**	19,10

Por último, en la muestra digital, como se lee en la Tabla 85, no se aprecian grandes distinciones en la producción léxica global entre hombres ($\overline{X}$ = 19,13) y mujeres ($\overline{X}$ = 19,25), pues la discrepancia entre ambos grupos es de 0,12 palabras, lo que podría apuntar a que, entre los dos sociolectos, hay más similitudes que diferencias. En efecto, cada grupo supera al otro en cuatro de las ocho áreas nocionales. En concreto, los actualizadores con los mayores promedios entre las féminas son *Comidas y bebidas* ($\overline{X}$ = 27,11), *La escuela* ($\overline{X}$ = 19,36), *La educación* ($\overline{X}$ = 17,57) y *El profesor*

($\overline{X}$ = 16,06). Por su parte, los hombres tienen los PP más altos en *Partes del cuerpo* ($\overline{X}$ = 28,06), *Juegos y distracciones* ($\overline{X}$ = 18,06), *La lectura* ($\overline{X}$ = 17,61) y *Habilidades docenes* ($\overline{X}$ = 12,56).

Tabla 85. Comparación de los $\overline{X}$ por sexo de LH2

Centro de Interés	Hombres	Mujeres
01. La lectura	17,61	17,41
02. El profesor	15,11	16,06
03. La educación	17,00	17,57
04. Juegos y distracciones	18,06	17,69
05. La escuela	18,44	19,36
06. Habilidades docentes	12,56	11,93
07. Partes del cuerpo	28,06	26,89
08. Comidas y bebidas	26,22	27,11
Media aritmética	19,13	**19,25**

En los Gráfico 17, 18 y 19, que muestran la distribución de los promedios de palabras por centros de interés en las muestras 1, 2 y 3, respectivamente, se observa que los tres conjuntos de datos presentan un patrón *grosso modo* similar respecto a la producción léxica de los grupos. Sin embargo, divergen en que, en primer lugar, los dos sociolectos del área de Pedagogía exponen una mayor distinción entre ellos, dejando casi separadas las líneas trazadas en la cuadrícula por los resultados, las cuales solo llegan a casi solaparse en los puntos referentes a los centros de interés 06, 04 y 07, lo que refleja la mayor distinción léxica entre hombres y mujeres en los actualizadores *Comidas y bebidas* y *La escuela*. En segundo lugar, los hombres y las mujeres de Letras Hispánicas, formato en papel, presentan un mayor grado de similitudes entre sus lexicones, pues las medias aritméticas por cada área nocional dibujan en la cuadrícula dos líneas paralelas bastante cercanas la una de la otra, mostrando mayor proximidad en los puntos concernientes a los CI 05, 04 y 06; pero mayor distanciamiento en *Partes de cuerpo*, en el que los chicos superan a las chicas. Y, en tercer lugar, los datos recolectados online reflejan la estrecha analogía existente entre los lexicones de los hombres y las mujeres, ya que las líneas trazadas, a partir de los valores de las medias, además de seguir el mismo modelo, llegan a solaparse casi por completo en todos los puntos de los centros de interés.

Gráfico 17. Distribución de los promedios de palabras por CI de EB

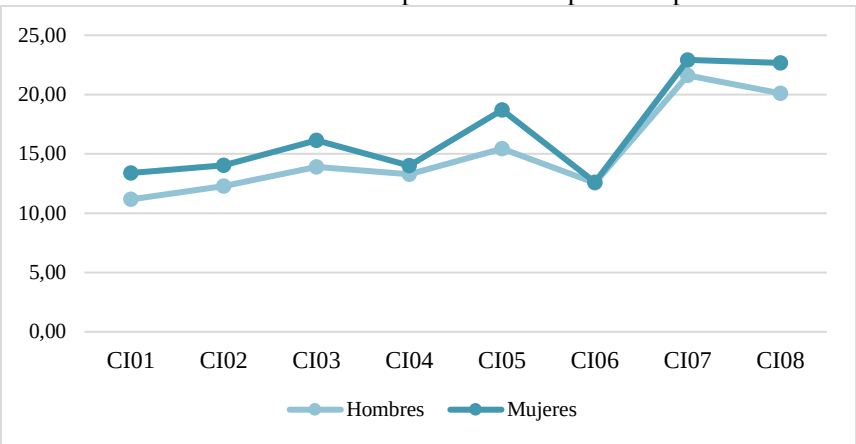


Gráfico 18. Distribución de los PP por CI de LH1

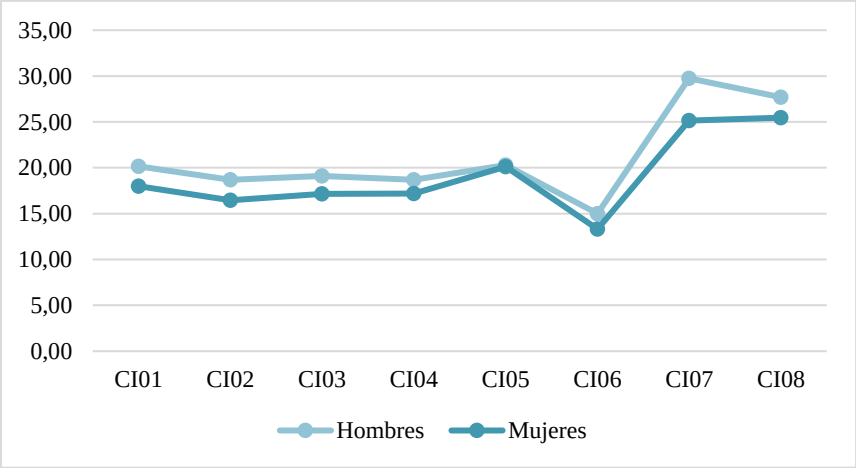
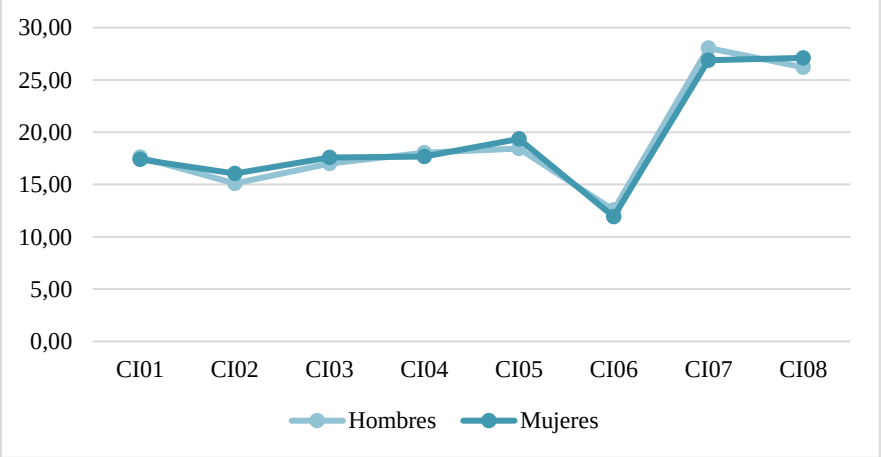


Gráfico 19. Distribución de los PP por CI de LH2



Por último, con el fin de determinar la significación de las diferencias de las medias de producción léxica entre hombres y mujeres de las tres muestras, se aplicó la prueba paramétrica t de Student a los tres conjuntos de datos; no sin antes calcular primero la homocedasticidad de cada grupo.

Esta se define a través de la prueba de Levene, que viene integrada al T-test de SPSS. Según los resultados, se asume la igualdad de varianzas de los promedios generales de las tres muestras, así como la de todas las áreas nocionales de los corpus de EB y LH1. Sin embargo, la distribución de las medias no es homogénea en las áreas nocionales: *El profesor* (sig. ,045) y *La educación* (sig. ,048) del corpus digital, pero sí lo es en el resto de los CI. En este contexto, se ha recurrido a la prueba no paramétrica U de Mann-Whitney para corroborar o refutar la H_0 en los datos de CI02 y CI03.

Tabla 86. Estadística de la variable Sexo en la muestra de EB

		Prueba de Levene para igualdad de varianzas		Prueba T para la igualdad de medias		
		F	Sig.	t	gl	Sig. (bilateral)
Número de palabras	Se han asumido varianzas iguales	,054	,816	-1,809	106	,073
	No se han asumido varianzas iguales			-1,789	24,037	,086
La escuela	Se han asumido varianzas iguales	,666	,416	-2,647	106	**,009**
	No se han asumido varianzas iguales			-2,444	22,671	,023

Tabla 87. Estadística de la variable Sexo en la muestra de LH1

		Prueba de Levene para igualdad de varianzas		Prueba t para la igualdad de medias		
		F	Sig.	t	gl	Sig. (bilateral)
Número de palabras	Se han asumido varianzas iguales	,060	,808	1,450	66	**,152**
	No se han asumido varianzas iguales			1,373	31,815	,179
Partes del cuerpo	Se han asumido varianzas iguales	,089	,766	2,006	66	**,049**
	No se han asumido varianzas iguales			1,778	28,182	,086

Tabla 88. Estadística de la variable Sexo en la muestra de LH2

		Prueba de Levene para igualdad de varianzas		Prueba t para la igualdad de medias		
		F	Sig.	t	gl	Sig. (bilateral)
Número de palabras	Se han asumido varianzas iguales	1,438	,234	-,084	86	**,933**
	No se han asumido varianzas iguales			-,096	32,222	,924

En las tablas 86, 87 y 88, se lee que el T-test arrojó que los índices de significación bilateral son superiores a alfa (,05) en las medias globales de las tres muestras. Específicamente, en el corpus de Educación Básica *p* es igual a ,073; en el de LH1, es ,152; y en el digital es ,933. En consecuencia, las diferencias de los promedios generales de palabras en relación con *Sexo* se deben al azar. No obstante, los análisis por cada eje temático de manera individual indica la existencia de asociaciones estadísticas significativa en los centros de interés *La escuela: muebles y materiales* (*p* = ,009), del corpus de Ed. Básica, como se detalla en la tabla 86[19]. Por su parte, en los datos de la muestra

[19] En las tablas de los análisis estadísticos bivariados –paramétricos y no paramétricos– se identifican, en negrita y sombreado, los resultados significativos.

tradicional de Letras Hispánicas se constata una relación estadísticamente significativa entre la variable y el actualizador *Partes del cuerpo* (p = ,049), como se ve en la Tabla 88. En tanto que, en la muestra digital, tanto el T-test como la prueba no paramétrica U de Mann-Whitney (Tabla 89) arrojaron valores de alfa por encima de ,05 en todos los CI analizados, por lo que se acepta la hipótesis de partida, según la cual el sexo de los sujetos no influye en el léxico disponible.

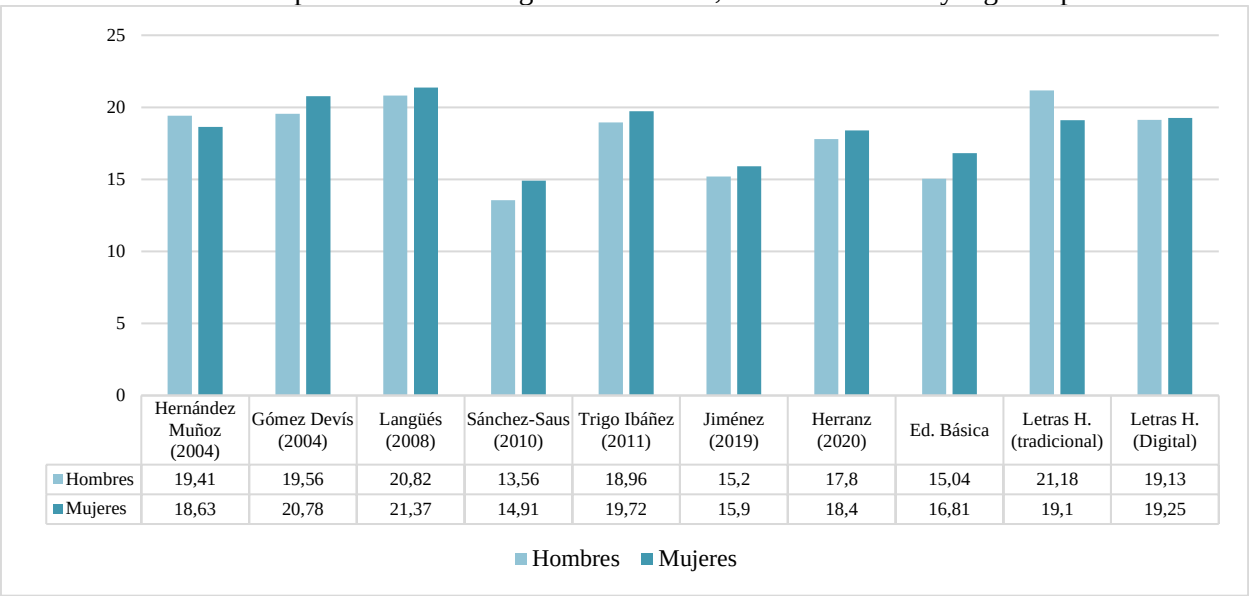
Tabla 89. Resultados de U de Mann-Whitney de LH2, según Sexo.

Tipo de muestra	*El profesor*	La educación
U de Mann-Whitney	597,000	607,500
W de Wilcoxon	768,000	778,500
Z	-,342	-,233
Sig. asintót. (bilateral)	**,732**	**,816**

En síntesis, estos resultados concernientes a la variable Sexo vienen a ratificar a grandes rasgos los hallazgos y afirmaciones declaradas en otras investigaciones, según las cuales este factor no presenta una incidencia totalmente clara en el caudal léxico de los informantes (Gómez Devís, 2004: 126). Por el contrario, de todas las categorías utilizadas en las diferentes pesquisas para describir el LD, esta pareciera ser la menos influyente, al menos en términos cuantitativos (Pacheco, 2016: 243). No obstante, en algunos contextos se aprecia que los hombres superan a las mujeres en cuanto a la producción léxica, como se lee en los resultados de Letras Hispánica, en formato papel, de esta tesis, así como en los trabajos de Hernández Muñoz (2006) y Pacheco (2016). Por el contrario, en otras pesquisas son las féminas las que tienen los promedios más altos, como se detalla en las muestras de Educación Básica y Letras Hispánicas (digital) de este estudio. En el gráfico 20, se comparan las medias aritméticas globales del LD en relación con el factor sexo entre esta y algunas investigaciones previas.

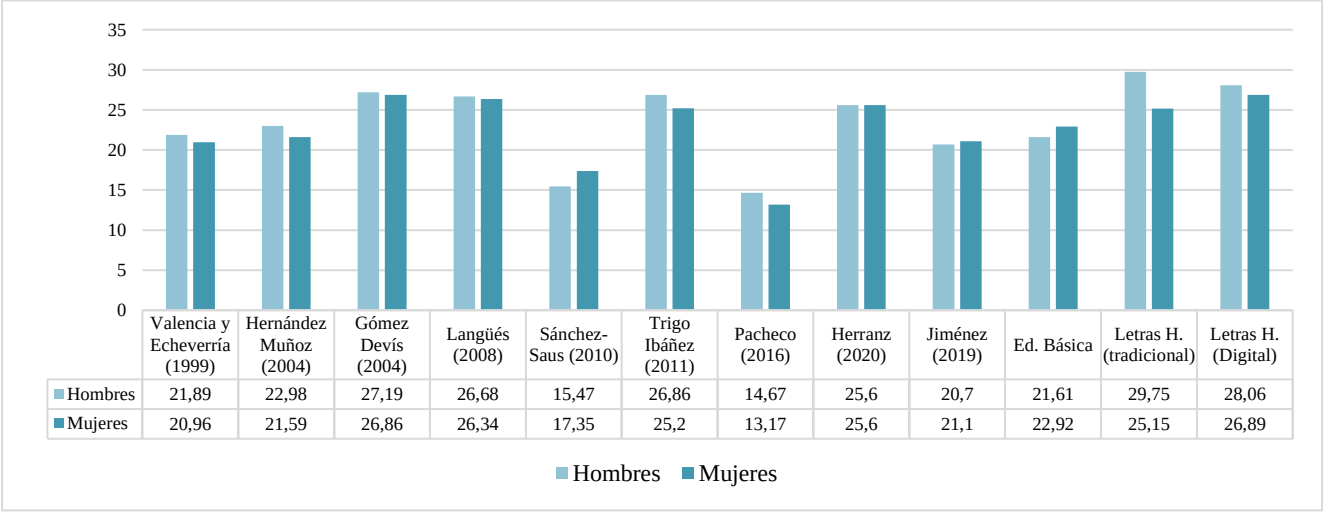
En esta misma línea, en algunos antecedentes se ha detallado que hay centros de interés en los que los hombres muestran promedios superiores a las mujeres y viceversa, lo que ha llevado a proponer la hipótesis de la incidencia del rol social en el léxico de los individuos. Sin embargo, al considerar los resultados de esta tesis, podría afirmarse que los hombres superan a las mujeres en el área nocional *Partes del cuerpo,* como se ha visto en Valencia y Echeverría (1999), Hernández Muñoz (2004), Gómez Devís (2004), Lagüéns (2008), Trigo Ibáñez (2011), Pacheco (2016), como se ilustra en el gráfico 20.

Gráfico 20. Comparación de los PP globales de *Sexo*, entre este estudio y algunos previos.



	Hernández Muñoz (2004)	Gómez Devís (2004)	Langüés (2008)	Sánchez-Saus (2010)	Trigo Ibáñez (2011)	Jiménez (2019)	Herranz (2020)	Ed. Básica	Letras H. (tradicional)	Letras H. (Digital)
Hombres	19,41	19,56	20,82	13,56	18,96	15,2	17,8	15,04	21,18	19,13
Mujeres	18,63	20,78	21,37	14,91	19,72	15,9	18,4	16,81	19,1	19,25

Gráfico 21. Comparación de los PP del CI *Partes del cuerpo*, según Sexo



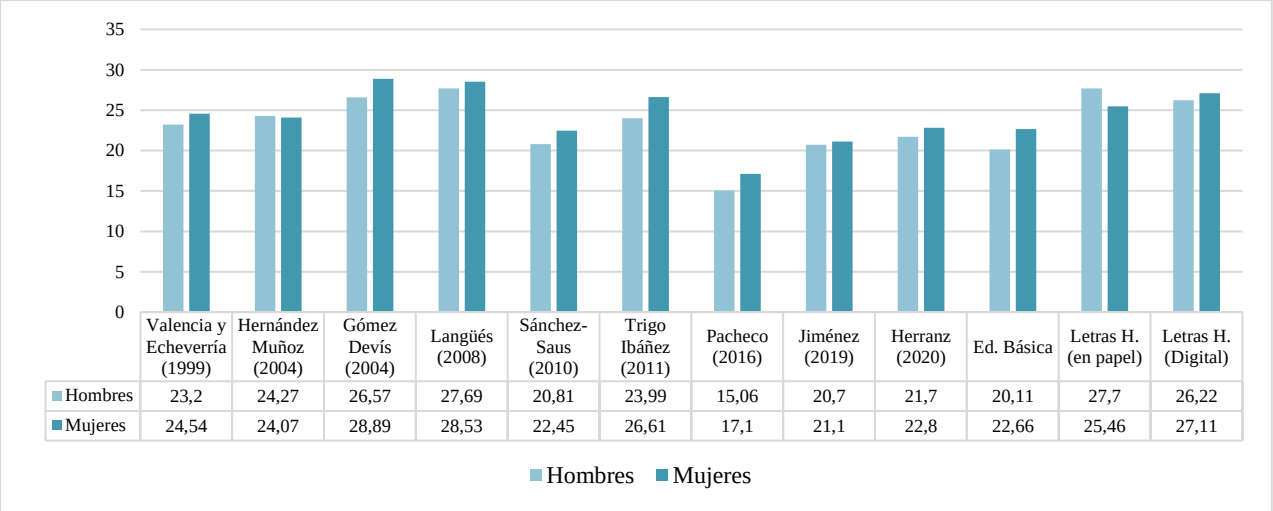
	Valencia y Echeverría (1999)	Hernández Muñoz (2004)	Gómez Devís (2004)	Langüés (2008)	Sánchez-Saus (2010)	Trigo Ibáñez (2011)	Pacheco (2016)	Herranz (2020)	Jiménez (2019)	Ed. Básica	Letras H. (tradicional)	Letras H. (Digital)
Hombres	21,89	22,98	27,19	26,68	15,47	26,86	14,67	25,6	20,7	21,61	29,75	28,06
Mujeres	20,96	21,59	26,86	26,34	17,35	25,2	13,17	25,6	21,1	22,92	25,15	26,89

No obstante, en los ejes temáticos *Comidas y bebidas* y *La escuela*, los resultados reflejan que la mayor productividad léxica se da en el sociolecto femenino. Esta tendencia se ve en los trabajos aquí comparados, salvo en los de Hernández Muñoz (2004) y la muestra de LH1, como se especifica en los gráficos 22 y 23.

Para cerrar este apartado, debe notarse de manera especial la comparación entre los resultados de esta tesis y los de Herranz (2020), respecto al CI *La educación*, ya que este solo había sido aplicado por la lingüista española, bajo el nombre (*Educación*, sin el artículo). Tanto en el grupo español como en el de Educación Básica chileno, el promedio de palabras de las mujeres supera al de los varones. Si bien este esquema es similar al de la muestra digital, la diferencia es ínfima. Y, contrariamente, la

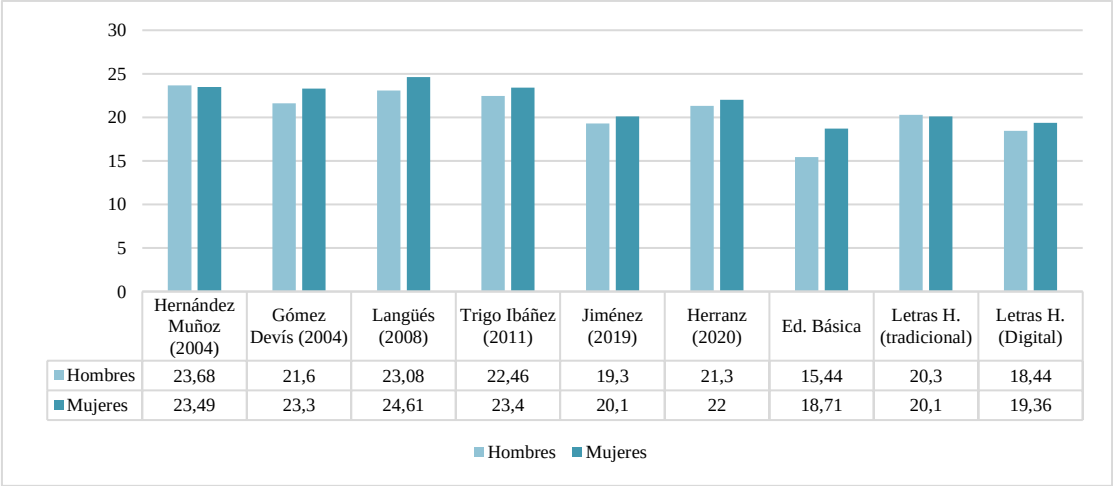
tendencia ascendente en el corpus (tradicional) de Letras la tienen los chicos ($\overline{X}$ = 19,1), llegando a superar por 2 puntos a sus congéneres europeos ($\overline{X}$ = 16,9), como se ilustra en el Gráfico 24.

Gráfico 22. Comparación de los PP de *Comidas y bebidas*, según Sexo



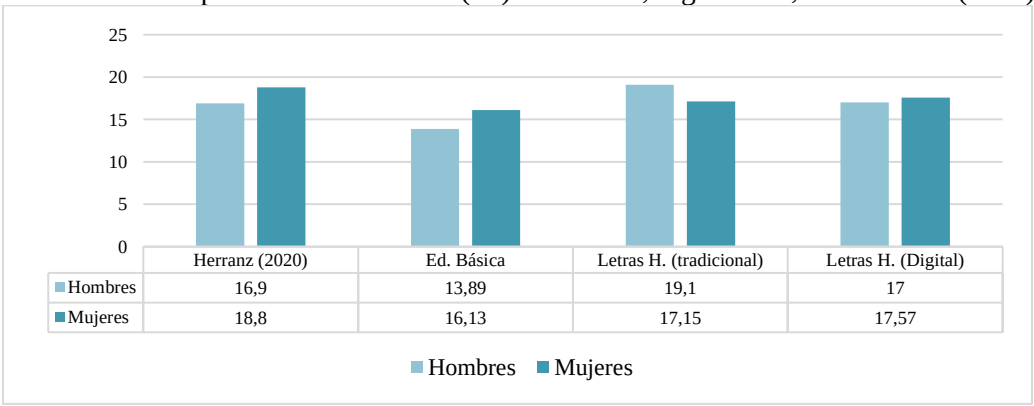
	Valencia y Echeverría (1999)	Hernández Muñoz (2004)	Gómez Devís (2004)	Langüés (2008)	Sánchez-Saus (2010)	Trigo Ibáñez (2011)	Pacheco (2016)	Jiménez (2019)	Herranz (2020)	Ed. Básica	Letras H. (en papel)	Letras H. (Digital)
Hombres	23,2	24,27	26,57	27,69	20,81	23,99	15,06	20,7	21,7	20,11	27,7	26,22
Mujeres	24,54	24,07	28,89	28,53	22,45	26,61	17,1	21,1	22,8	22,66	25,46	27,11

Gráfico 23. Comparación de los PP de *La escuela*, según Sexo



	Hernández Muñoz (2004)	Gómez Devís (2004)	Langüés (2008)	Trigo Ibáñez (2011)	Jiménez (2019)	Herranz (2020)	Ed. Básica	Letras H. (tradicional)	Letras H. (Digital)
Hombres	23,68	21,6	23,08	22,46	19,3	21,3	15,44	20,3	18,44
Mujeres	23,49	23,3	24,61	23,4	20,1	22	18,71	20,1	19,36

Gráfico 24. Comparación de los PP de (La) Educación, según Sexo, con Herranz (2020)



	Herranz (2020)	Ed. Básica	Letras H. (tradicional)	Letras H. (Digital)
Hombres	16,9	13,89	19,1	17
Mujeres	18,8	16,13	17,15	17,57

4.3.2. Carrera

La revisión bibliográfica da cuenta de las escasas pesquisas en el ámbito de la DL en las que se hayan contrastados los lexicones de estudiantes de distintas carreras; de hecho, pueden mencionarse, únicamente hasta ahora, los trabajos de Guerra y Gómez (2004), Blanco *et al.* (2020), Herranz (2020), Herranz y Marcos-Calvo (2021). En este sentido, puede afirmarse que esta tesis busca también contribuir con los estudios comparativos del léxico potencial de diversas carreras y/o comunidades discursivas (Parodi, 2004; Rodríguez Romero, 1998). Así pues, en este apartado, se analizan las relaciones estadísticas entre el vocabulario disponible y la variable *Carrera*; en particular, los lexicones de los alumnos de Educación Básica y Letras Hispánicas que realizaron las pruebas en papel. Se parte del supuesto de que, aunque los promedios globales no presentarán grandes diferencias entre los programas de estudios; sin embargo, habría la posibilidad de que las medias fueran mayores en algunos centros de interés a razón de considerarse más ligados a las mallas curriculares. En virtud de los cálculos estadísticos, se asume la hipótesis (H_0) de que el factor *Carrera* no tiene ninguna incidencia en el léxico disponible de los encuestados, en general ni por área nocional.

Al comparan los datos de producción léxica de los participantes del subcorpus 1, se observa que el valor más alto de la media aritmética general lo tiene el grupo de Letras Hispánicas ($\overline{X}$ = 19,71) versus el de Educación Básica ($\overline{X}$ = 16,51), presentando una diferencia de 3,2 palabras. En este mismo orden de idea, la muestra del área de Humanidades también expone los mayores promedios en todos los centros de interés, los cuales sobresalen en *Partes del cuerpo* ($\overline{X}$ = 26,50), *Comidas y bebidas* ($\overline{X}$ = 26,12) y *La escuela* ($\overline{X}$ = 20,16), rebasando la media global, como se detalla en Tabla 90.

Tabla 90. Comparación de los PP entre EB y LH1

Centros de interés	Ed. Básica	Letras H.
01. La lectura	13,01	18,62
02. El profesor	13,75	17,12
03. La educación	15,76	17,72
04. Juegos y distracciones	13,88	17,63
05. La escuela	18,17	20,16
06. Habilidades docentes	12,60	13,81
07. Partes del cuerpo	22,70	26,50
08. Comidas y bebidas	22,23	26,12
Media aritmética	16,51	**19,71**

Estos resultados dejan ver las diferencias existentes en la productividad léxica de los dos grupos comparados, que al graficarlos trazan dos líneas paralelas en la cuadrícula, una sobre la otra, con una separación evidente. Esta se marca aún más en los puntos que simbolizan los centros de interés

La lectura –alcanzando una diferencia de 5,61 puntos en las medias aritméticas–, *Comidas y bebidas* (3,89 puntos), *Partes del cuerpo* (3,8) y *Juegos y distracciones* (3,75 puntos). Opuestamente, los grupos llegan casi a solaparse en el CI06. *Habilidades docentes*, donde la separación es de apenas 1,21 puntos de promedio. Estos datos puedes verse en el Gráfico 25.

Gráfico 25. Contraste de los PP entre EB y LH1

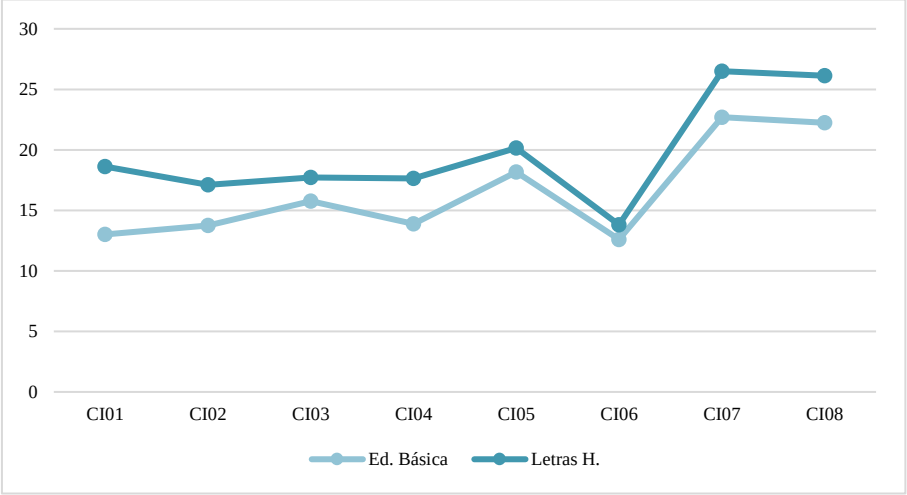


Tabla 91. Resultados de t de Student, según *Carrera*.

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias		
		F	Sig.	t	gl	Sig. (bilateral)
La lectura	Se han asumido varianzas iguales	3,288	,072	-6,688	174	,000
	No se han asumido varianzas iguales			-6,417	123,545	,000
La escuela	Se han asumido varianzas iguales	2,247	,136	-2,368	174	,019
	No se han asumido varianzas iguales			-2,248	118,867	,026
Comidas y bebidas	Se han asumido varianzas iguales	1,550	,215	-4,206	174	,000
	No se han asumido varianzas iguales			-4,094	129,995	,000

Tabla 92. Resultados de U de Mann-Whitney respecto al factor *Carrera*.

Pruebas no paramétricas	Palabras totales	El profesor	Juegos y distracciones	Partes del cuerpo
U de Mann-Whitney	2363,500	2452,500	2353,500	2485,000
W de Wilcoxon	8249,500	8338,500	8239,500	8371,000
Z	-3,976	-3,713	-4,014	-3,613
Sig. asintót. (bilateral)	**,000**	**,000**	**,000**	**,000**

En cuanto a los resultados estadísticos, la prueba de Levene indicó que no existe homocedasticidad entre las medias de los grupos de *Total general*, *El profesor*, *Juegos y distracciones*

y *Partes del cuerpo*, de manera que estos fueron analizados con el test no paramétrico U de Mann-Whitney. Este arrojó un valor de significación asintótica bilateral igual a ,000 para el promedio global, así como para los CI02, CI04 y CI07 (Tabla 92). En cuanto a los demás actualizadores, se les aplicó la prueba t de Student, ya que se asumía la igualdad de varianzas, según el examen de Levene. El T-test señaló que los ejes temáticos *La lectura* y *Comidas y bebidas*, ambos con una significación bilateral de ,000; y *La escuela*, con alfa igual a ,019, tienen significación estadística respecto a la variable *Carrera*, como se lee en la Tabla 91. En este contexto, entonces, cabe afirmar que se rechaza la hipótesis de partida (H_0) y se acepta la alternativa: la carrera de los participantes sí influye en el léxico disponible relacionado con siete de los ochos CI de esta tesis, salvo en los CI03 y CI06.

En resumen, con base en estos resultados, puede señalarse que el factor *Carrera* parece motivar la producción léxica de los participantes, en seis de los ochos ejes temáticos bajo análisis. Así pues, puede afirmarse que los estudiantes de Letras Hispánicas tienen una mayor producción léxica que sus compañeros de Educación Básica, ya que los análisis estadísticos bivariados han arrojado valores de significación tanto al nivel ,05 como al ,001. En este contexto, debe resaltarse que los actualizadores *Partes del cuerpo* (p = ,000), *Comidas y bebidas* (p = ,000) y *La escuela* (p = ,019), además de ser los más productivos, resultaron también estadísticamente significativos en relación con esta variable, en virtud de lo cual, podría afirmarse que apuntar los alumnos de Letras tienen un mayor dominio léxico. Sin embargo, debe señalarse que estos tres centros de interés, junto con *Juegos y distracciones* (p = ,000), competen a categorías semánticas de conocimiento general, que, a grandes rasgos, suelen ser más productivos que los campos nocionales de especialidad, como argumentan Quintanilla y Salcedo (2019).

En atención a lo anterior, se consideran más relevantes aún los resultados de las áreas temáticas *La lectura* y *El profesor*, ya que pueden considerarse como nombres que evocan campos semánticos vinculados más directamente con los saberes de ambos programas universitarios; específicamente, pareciera que el CI01 se inclina más hacia Letras, mientras que el CI02, a Educación. Las pruebas estadísticas paramétricas y no paramétricas llevan a rechazar la hipótesis de partida y aceptar la alternativa, según la cual la carrera de los sujetos sí influye en el léxico disponible de los encuestados (cf. Guerra y Gómez, 2003). Así pues, se planteó el supuesto de que el caudal léxico de los alumnos de Letras sería más productivo en el CI01. *La lectura*, mientras que el de Educación lo sería en el CI02. *El profesor*. El primer supuesto quedó demostrado, tanto por el promedio de palabras como por el T-test; de hecho, este último arrojó un valor de p = ,000 en el CI01 en favor del grupo de Letras. Sin embargo, no ocurrió lo mismo con el segundo supuesto, debido a que, además de la media

aritmética, el cálculo de U de Mann-Whitney mostró un *p* = ,000 en el CI02 en los datos de Letras, mas no así en los de Educación.

Para cerrar esta sección, se contrastan las medias aritméticas de los CI *La educación* y *La escuela* de esta investigación con las de Herranz (2020). En la Tabla 93 se observa que los promedios de palabras de los universitarios españoles de Educación Primaria y Educación Infantil superan a los de los grupos chilenos, tanto en el CI03 como en el CI05. De manera particular, los resultados del actualizador *La educación* presentan medias bastante similares entre los datos españoles y Letras Hispánicas, encontrando una distinción de solamente 0,71 puntos promediales de palabras; pero hay una diferencia de 2,72 palabras con la muestra de Educación Básica. En cuanto al CI *La escuela*, la comparación de los promedios del grupo chileno de Educación Básica con el español refleja una diferencia de 3,84 puntos; pero esta distancia de medias se achica cuando se contrastan los datos de Españas con la muestra de Letras, ya que entre ambas hay una disparidad de 1,79 puntos.

Tabla 93. Comparación del presente estudio con el de Herranz (2020)

CI	Presente estudio				Herranz (2020)			
	Ed. Básica		Letras H.		Ed. Primaria		Ed. Infantil	
	NP	PP	NP	PP	NP	PP	NP	PP
La educación	1693	15,68	1203	17,69	5588	**18,3**	5270	**18,4**
La escuela	1961	18,16	1374	20,21	6724	**22**	6189	**21,6**

4.3.3. Año de curso

En este apartado se describen los resultados de la producción léxica de los universitarios de las Facultades de Educación y Letras respecto al factor Año de curso.

Tabla 94. Promedio de palabras de la muestra de EB según Año de curso

Ed. Básica	1.er. Año	4.° Año
01. La lectura	13,23	12,28
02. El profesor	14,31	11,88
03. La educación	16,16	14,44
04. Juegos y distracciones	14,08	13,20
05. La escuela	18,67	16,48
06. Habilidades docentes	12,60	12,60
07. Partes del cuerpo	23,11	21,36
08. Comidas y bebidas	22,72	20,60
Media aritmética	**16,86**	15,36

En la tabla 94, correspondiente a los datos de Educación Básica, se observa que la variante 1.er Año ostenta el valor más alto de la media aritmética del grupo, con 16,81 puntos; mientras que la variante 4. ° Año refleja un $\overline{X}$= 15,36; es decir, tienen una diferencia de 1,5 palabras. Del análisis por área nocional, el grupo del primer nivel de escolaridad universitaria supera al de cuarto en siete de los

ocho actualizadores, salvo en *Habilidades y cualidades docentes*, en el que ambos tienen el mismo promedio, a saber: 12,60 palabras. Los ejes temáticos que sobrepasan los promedios de cada sociolecto son *Partes del cuerpo*, *Comidas y bebidas* y La escuela.

Estos resultados se ilustran en el gráfico 26, en el que puede apreciarse un patrón casi similar en las líneas de los promedios de palabras por centros de interés entre el sociolecto de 1.er. Año y el de 4° Año de EB. La oscilación de las medias por actualizador es baja entre uno y otro grupo de la muestra, al punto de que se solapan en el CI06. Pero se detalla una leve separación respecto al CI02 (las tendencias se mueven en direcciones distintas), puesto que el PP aumenta en los listados del colectivo de recién ingreso; contrariamente, disminuye en los datos de quienes están cercanos a egresar. Igualmente, el gráfico deja manifiesto cuáles son los actualizadores que sobresalen gracias a su productividad léxica.

Gráfico 26. Distribución del PP por CI de la muestra de Ed. Básica, según año de curso

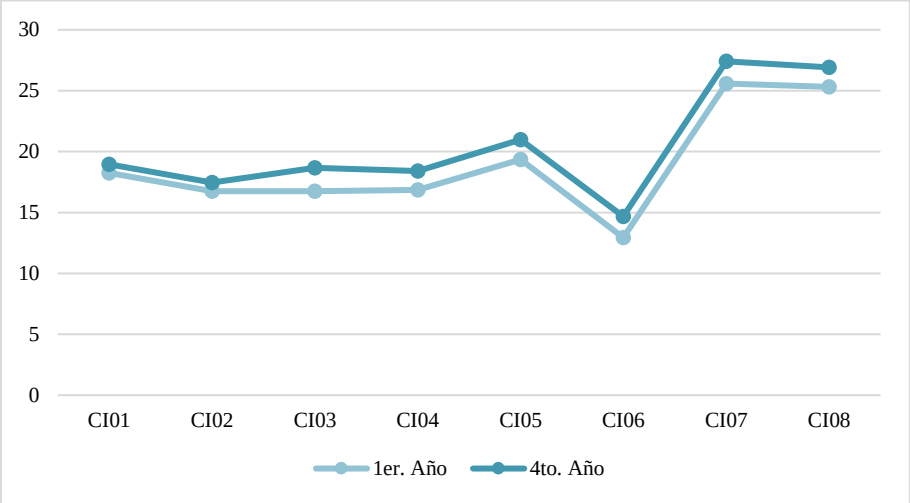
En relación con la muestra de Letras Hispánicas recolectada de manera tradicional, los resultados indican que el sociolecto de 4° Año tiene una media aritmética general de 20,44 palabras, 1,46 puntos por encima del sociolecto de 1.er. Año, que logró un promedio de 18,98 unidades léxicas en total. Igualmente, los listados de quienes se acercan al final de la carrera muestran los valores más altos en los ochos CI del estudio, en contraste con los de recién ingreso al programa universitario. Sobre los PP por subgrupo, los actualizadores que superan la media global de 1.er Año son Partes del cuerpo ($\overline{X}$= 25,59), *Comidas y bebidas* ($\overline{X}$= 25,32), y *La escuela: muebles y materiales* ($\overline{X}$= 19,35). Estos tres ejes temáticos también son los más productivos del grupo de 4. ° Año, exactamente: CI07 ($\overline{X}$= 27,41), CI08 ($\overline{X}$= 26,91) y CI05 ($\overline{X}$= 20,97), como se lee en la tabla 95.

Tabla 95. Promedio de palabras de LH1, según año de curso

Centros de interés	1.er Año	4.° Año
01. La lectura	18,26	18,97
02. El profesor	16,76	17,47
03. La educación	16,76	18,68
04. Juegos y distracciones	16,85	18,41
05. La escuela	19,35	20,97
06. Habilidades docentes	12,94	14,68
07. Partes del cuerpo	25,59	27,41
08. Comidas y bebidas	25,32	26,91
Media aritmética	18,98	**20,44**

La distribución de las medias de palabras por actualizador traza una figura que expone las semejanzas del caudal léxico de los sociolectos de 1.er y 4. ° Año de Letras Hispánicas, ya que las líneas que representan a cada subgrupo se posicionan paralelamente, dibujando la misma forma. En esta se evidencia que la mayor cercanía se da en los CI01 y CI02, en cuyos casos los vectores se solapan medianamente; mientras que, por el contrario, se aprecia un distanciamiento entre las variantes en los ejes temáticos 03 y 04. Igualmente, se detallan las áreas nocionales más y menos productivas, ya que los puntos que las identifican resaltan en el gráfico 27.

Gráfico 27. PP por CI de la muestra de LH1, según año de curso



En la muestra recogida de forma digital, los estudiantes de LH de 1.er Año presentan la media aritmética general más alta versus a los de 4. ° Año, con resultados equivalentes a 20,12 y 18,51 palabras, respectivamente. Esto representa una discrepancia de 1,61 puntos entre ambos sociolectos. En consonancia con lo anterior, el grupo de alumnos de recién ingreso también supera al sociolecto avanzado en cuanto a los promedios de piezas léxicas por área nocional, de los que únicamente *Comidas y bebidas* ($\overline{X}$= 29,46) y Partes del cuerpo ($\overline{X}$= 28,64) sobrepasan la media general de 1.er

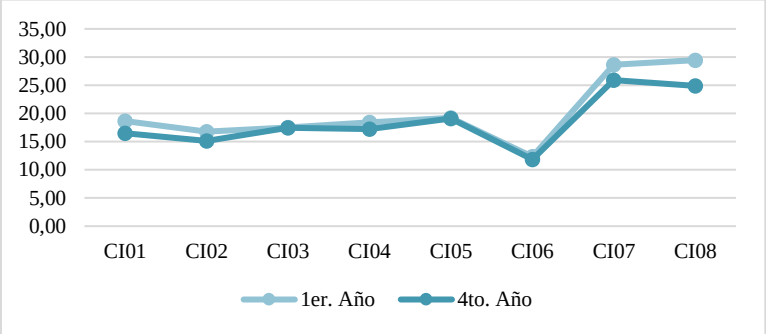
Año. Por su parte, en el sociolecto de 4. ° Año son tres los actualizadores que pasan el límite del promedio general, específicamente: CI07 ($\overline{X}$= 25,92), CI08 ($\overline{X}$= 24,92) y CI05 ($\overline{X}$= 19,12); en cambio, en los de 1.er Año son solo dos: CI08 y CI07, como se expone en la tabla 96.

Tabla 96. Promedio de palabras de la muestra digital, según año de curso

Centros de interés	1.er Año	4.° Año
01. La lectura	18,64	16,51
02. El profesor	16,77	15,14
03. La educación	17,49	17,43
04. Juegos y distracciones	18,44	17,22
05. La escuela	19,23	19,12
06. Habilidades docentes	12,33	11,84
07. Partes del cuerpo	28,64	25,92
08. Comidas y bebidas	29,46	24,92
Media aritmética	**20,12**	18,51

Si se detallan estos datos, puede apreciarse una oscilación bastante baja entre los promedios de palabras por área nocional, ya que trazan líneas contiguas semejantes y muy cercanas entre los dos sociolectos. En consecuencia, los puntos de los actualizadores La educación y La escuela se superponen casi por completo, efecto que se evidencia también, aunque en menor medida, en el CI06. Opuestamente, la tendencia del eje temático 08 es dispar, ya que las líneas se dirigen a puntos contrarios en los dos sociolectos, como se ilustra en el gráfico 28, a continuación.

Gráfico 28. Distribución del PP por CI de la muestra digital, sobre año de curso



Con el fin de determinar la hipótesis sobre la posible incidencia de los niveles de escolaridad en el LD de los encuestados, se ha realizado la prueba paramétrica t de Student, no sin antes verificar la homocedasticidad del corpus. En consecuencia, la prueba de Levene arrojó que todos los datos –tanto en la muestra de EB como en la de LH (digital)– se acepta la H_0, concerniente a la igualdad de varianzas. Por su parte, en la muestra de LH (tradicional), se asume la homogeneidad en el Número de palabras y en los actualizadores 02, 03, 04, 05, 06 y 07. Sin embargo, este criterio no se cumple en

las áreas nocionales 01 y 08, cuya significación es ,029 y, 015, respectivamente; por lo que se asume que tienen varianzas disímiles. Por ende, no es viable interpretar los resultados de t de Student de estas dos áreas nocionales, pero sí los de la prueba no paramétrica U de Mann-Whitney.

Sobre los cálculos del T-test, los resultados indican que no hay una asociación estadística significativa entre el factor Año de la carrera y el Número de palabras de ninguna de las tres muestras, ya que, en todas, el valor de significación bilateral supera el límite de error Tipo I (,05). Particularmente, el conjunto de Educación Básica alcanzó un alfa igual a ,084; mientras que el de LH1 es ,272; y el de LH2 equivale a ,162. En conformidad a estos, puede afirmarse que las diferencias de los promedios globales respecto al año de escolaridad se deben al azar.

A la par de estos cómputos, la prueba paramétrica plantea una relación estadística significativa entre el nivel de escolaridad universitaria y el caudal léxico de los encuestados de Educación Básica en las áreas nocionales *el profesor* y *la escuela: muebles y materiales*. En virtud de que estos dos actualizadores presentan un α igual a ,022 y, 050, respectivamente, como se lee en la tabla 97.

Tabla 97. Resultados del t-test de EB, respecto a Año de curso

		F	Sig.	t	gl	Sig. (bilateral)
El profesor	Se han asumido varianzas iguales	,097	,756	2,320	106	**,022**
	No se han asumido varianzas iguales			2,274	38,434	,029
La escuela	Se han asumido varianzas iguales	,735	,393	1,985	106	**,050**
	No se han asumido varianzas iguales			1,867	36,298	,070

Contrariamente, los resultados concernientes a la muestra en papel de los alumnos de Letras Hispánicas sugieren que no existe una relación estadística entre el LD y el nivel de escolaridad universitaria, ya que en todos los ejes temáticos se aprecian valores de alfa superiores a ,05. De modo semejante, la prueba no paramétrica U de Mann-Whitney apunta a que, en los casos de los CI La lectura (p = ,931) y *Comidas y bebidas* (p = ,291), tampoco hay una asociación entre las variables dependientes y la independiente, puesto que los valores de p sobrepasan el límite de error. Por ende, la variación observada en estos casos está sujeta al azar.

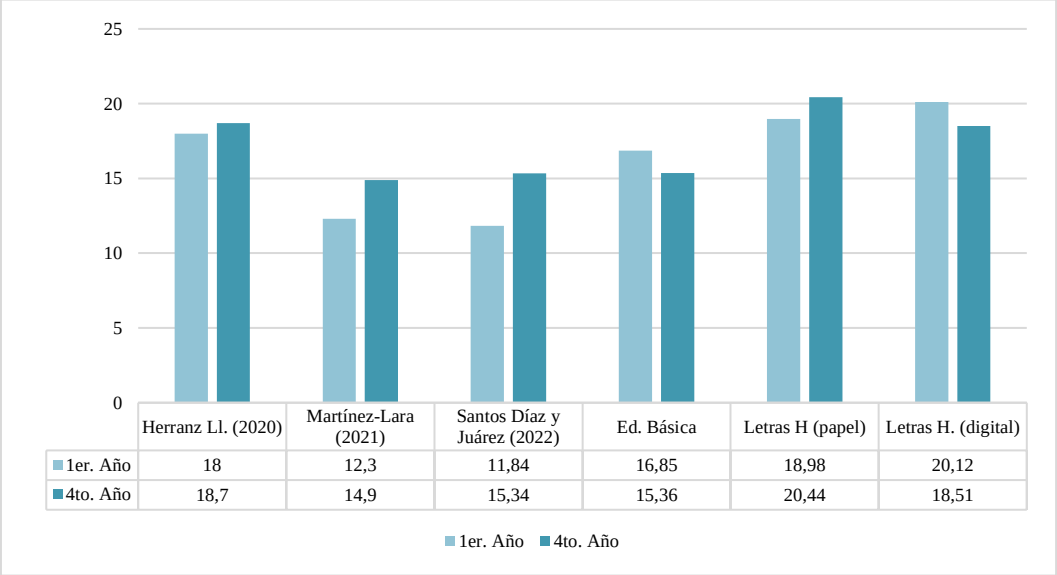
Sobre la muestra digital, según el t-test, puede señalarse que se rechaza la H_0 únicamente del centro de interés *comidas y bebidas*, en virtud de que dicho conjunto expone una significación bilateral igual a ,016. Entonces, puede afirmarse que, para este caso, los datos no son azarosos, por lo que el factor Año de curso incide en el léxico referido al tema de la ingesta de alimentos y bebidas en el conjunto de alumnos de LH (digital), como se ve en la tabla siguiente.

Tabla 98. Resultados del t-test de la muestra digital, sobre año de curso

		F	Sig.	t	gl	Sig. (bilateral)
Comidas y bebidas	Se han asumido varianzas iguales	1,176	,281	2,461	86	**,016**
	No se han asumido varianzas iguales			2,431	77,167	,017

Finalmente, puede indicarse que los resultados hallados en este estudio parecen contradecir lo reportado por la mayoría de las investigaciones previas –salvando el caso de la muestra en papel de Letras Hispánicas–. Esto debido a que los antecedentes han demostrado que los estudiantes de mayor nivel de escolaridad tienden a exponer los índices léxicos más elevados que sus congéneres de los niveles inferiores (cf. Echeverría *et al.* 1987; Salcedo y Leo 2013; Ferreira *et al.* 2014; Jiménez, 2019; Herranz, 2020; Martínez-Lara, 2021; Zhou, 2021; Santos Díaz y Juárez, 2022, entre otros). En el gráfico 29, seguidamente, se ilustran los resultados del factor Año de la carrera de algunos trabajos cuyos sujetos tienen características extralingüísticas similares a los de esta tesis.

Gráfico 29. Comparación intermuestral de los PP, según la variable año de curso



	Herranz Ll. (2020)	Martínez-Lara (2021)	Santos Díaz y Juárez (2022)	Ed. Básica	Letras H (papel)	Letras H. (digital)
1er. Año	18	12,3	11,84	16,85	18,98	20,12
4to. Año	18,7	14,9	15,34	15,36	20,44	18,51

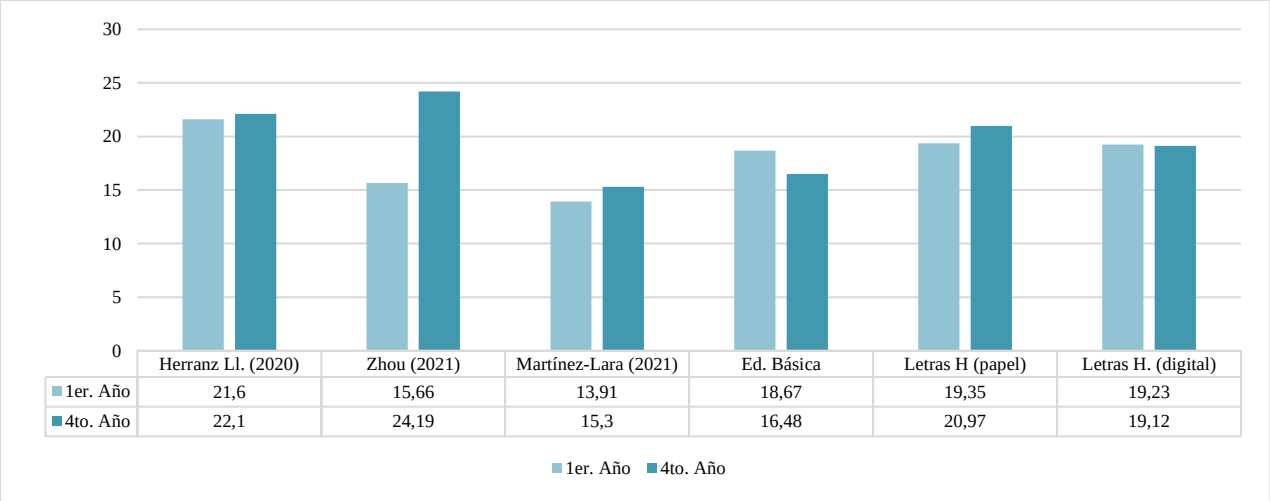
Tanto Herranz (2020) como Martínez-Lara (2021) y Santos Díaz y Juárez (2022) han analizado el caudal léxico de estudiantes universitarios de carreras del área de Pedagogía en relación con variables extralingüísticas, entre las que se halla año de escolaridad. En dichas pesquisas, así como en la muestra tradicional de Letras Hispánicas se observa que los sujetos avanzados aventajan a los de los primeros niveles, respecto a los promedios de palabras globales. En este sentido, puede indicarse que –al menos para el conjunto de LH (técnica en papel)– parece haber una relación entre el nivel de escolaridad y el caudal léxico de los individuos. De estos datos, debe resaltarse que –como se aprecia en el

subapartado previo– el conjunto de LH expone índices más altos que los expuestos en los tres antecedentes nombrados al inicio del párrafo. En consecuencia, este contraste apunta a que los alumnos del programa de Letras tienden a ser léxicamente más productivos que los del área de Pedagogía, tanto de España como de Chile.

Opuestamente, los resultados de las muestras de EB y LH2 del presente trabajo muestran una mayor productividad léxica entre los sociolectos de 1.er Año versus los de 4.o Año, lo que parece no ser coherente con los estudios previos y, por ende, van en dirección contraria respecto a la hipótesis planteada. Sin embargo, estos números podrían deberse a razones externas a la investigación y de índole personal, las cuales se escapan de la observación directa de los analistas. No obstante, debe resaltarse que, a la luz de los cálculos matemáticos, estos resultados son estadísticamente poco fiables. A pesar de lo anterior, como se aprecia en los cálculos del t-test, dos de los ochos centros de interés resultaron significativos, a saber: *La escuela: muebles y materiales* y *Comidas y bebidas*. A razón de los cual, en las subsiguientes líneas se contrastan los resultados de dichas áreas nocionales con las de los trabajos de Herranz (2020), Martínez-Lara (2021) y Zhou (2021), donde se analizaron dichos actualizadores entre el alumnado de pedagogía.

En el caso del CI05, los datos de la muestra de Educación Básica resultaron estadísticamente significativa a favor de 1.er Año, con p = ,050. En relación con los demás trabajos, los PP de la muestra de pedagogía son mayores que los del grupo de universitarios chilenos analizados por Martínez-Lara (2021), en cuyo caso la variable también es significativa al nivel ,001 entre los discentes avanzados. En cuanto al alumnado español (Herranz, 2020; Zhou, 2021), estos presentan los promedios más altos entre los conjuntos de niveles superiores, con valores muy por encima de la muestra chilena de educación aquí estudiada, siendo más evidente entre los datos de Zhou (2021), cuya diferencia entre los dos sociolectos es casi de 10 puntos, concretamente: 8,53 lemas, como se ilustra en el gráfico 30, a continuación.

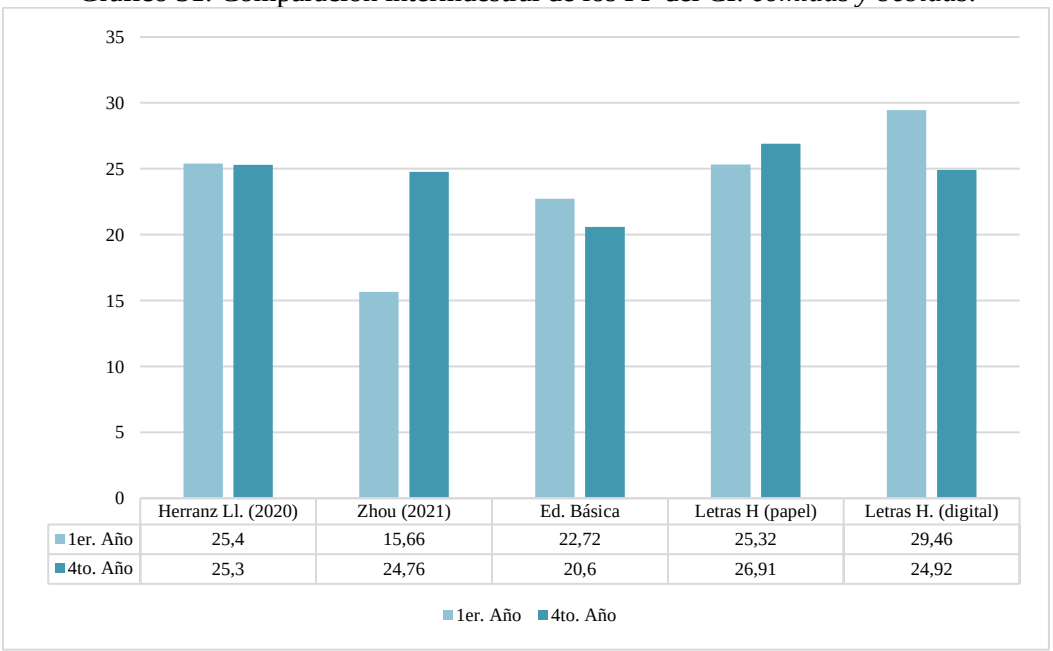
Gráfico 30. Contraste intermuestral de los PP del CI *La escuela*



	Herranz Ll. (2020)	Zhou (2021)	Martínez-Lara (2021)	Ed. Básica	Letras H (papel)	Letras H. (digital)
1er. Año	21,6	15,66	13,91	18,67	19,35	19,23
4to. Año	22,1	24,19	15,3	16,48	20,97	19,12

Sobre el CI08, la prueba paramétrica t de Student arrojó una significación de p = ,016 para el 1.er Año en la muestra digital de Letras Hispánicas, cuyo valor supera al resto de las investigaciones aquí comparadas. De manera detallada, se aprecia que entre ambos cursos de LH la distinción es clara, logrando una separación de 4,54 palabras. Esta es superada únicamente por la pesquisa de Zhou (2021), cuya disimilitud entre los niveles de escolaridad universitaria alcanza 9,1 lemas. Sin embargo, las diferencias se contraen en el análisis de Herranz (2020), en cuyo caso el mayor $\overline{X}$ lo tiene el subgrupo de 1.er Año, pero por apenas 0,1 lema, según se aprecia en el gráfico 31, seguidamente.

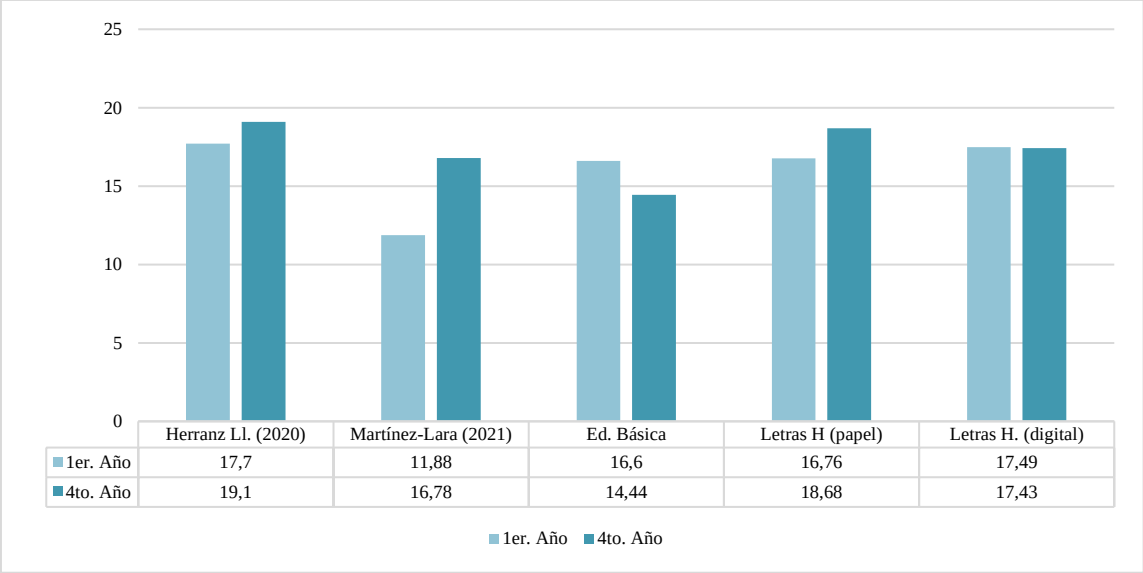
Gráfico 31. Comparación intermuestral de los PP del CI: *comidas y bebidas*.



	Herranz Ll. (2020)	Zhou (2021)	Ed. Básica	Letras H (papel)	Letras H. (digital)
1er. Año	25,4	15,66	22,72	25,32	29,46
4to. Año	25,3	24,76	20,6	26,91	24,92

Por último, se han querido contrastar los resultados de los PP del actualizador La educación, ya que es uno de los novedosos aplicado en investigaciones recientes; específicamente, lo inauguró Herranz (2018). Según se lee en el gráfico 37, este CI ha sido más productivo en el trabajo desarrollado en España, donde la variante 4.º Año obtuvo el $\overline{X}$ de palabras más alto respecto al resto de las muestras, a saber: 19,1 vocablos, con una diferencia de 1,4 puntos. Este esquema se repite en el conjunto de Letras Hispánicas (tradicional), cuya separación a favor del 4.º Año es de 1,92 lexías. Pero la mayor distinción resalta en la investigación de Martínez-Lara (2021), en la que los alumnos avanzados tienen 4,9 lexemas por encima de los de iniciación. No obstante, este patrón se rompe con el estudiantado chileno del área de Pedagogía, en cuyo caso el grupo del nivel inicial es el que expone la media aritmética más elevada ($\overline{X}$ = 16,60) frente al del último escalafón ($\overline{X}$ = 14,44), con una separación de 2,16 palabras, según se detalla en el gráfico 32.

Gráfico 32. Contraste intermuestral de los PP del CI *La educación*



	Herranz Ll. (2020)	Martínez-Lara (2021)	Ed. Básica	Letras H (papel)	Letras H. (digital)
1er. Año	17,7	11,88	16,6	16,76	17,49
4to. Año	19,1	16,78	14,44	18,68	17,43

4.3.4. Formato de las pruebas

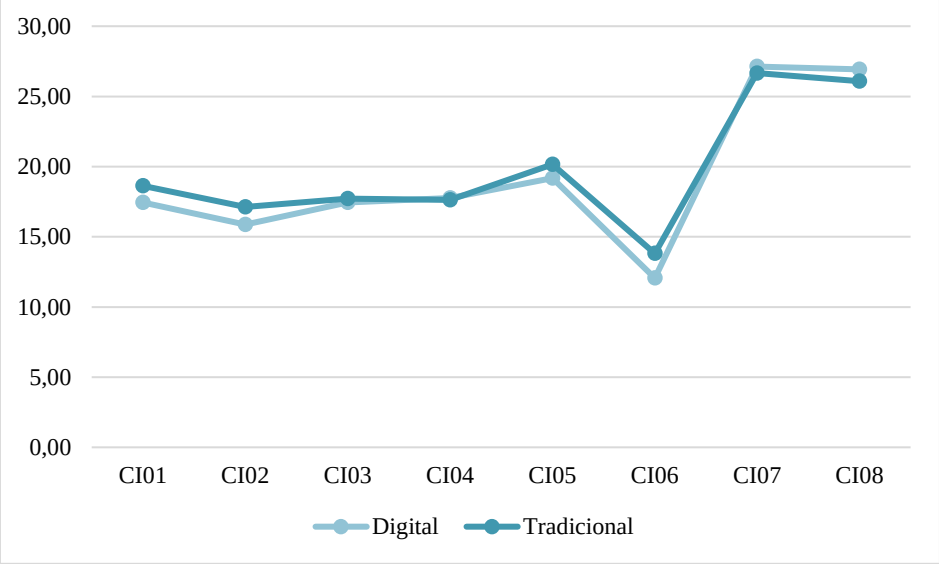
En el subapartado 3.5.3., se ha descrito y contrastado la producción léxica de los estudiantes de Letras Hispánicas que contestaron las pruebas de disponibilidad léxica tanto en papel (método tradicional) como en la página web (método digital). Debido a lo anterior, en este apartado, se ahondará respecto a los resultados estadísticos bivariados entre el Tipo de método y el caudal léxico de los encuestados, con el fin de abordar uno de los objetivos principales de esta tesis. Recapitulando, en el epígrafe X se evidenció que la media general de la muestra tradicional es mayor que la de la digital, 19,71 *vs* 19,23 piezas léxicas, respectivamente. No obstante, los valores de ambos corpus son fuertemente similares, con una distinción de apenas 0,48 lexías, como se aprecia en la tabla 99.

Tabla 99. Comparación del PP de los alumnos de Letras H., según formato de prueba

CI	Digital	Tradicional
01. La lectura	17,45	18,62
02. El profesor	15,85	17,12
03. La educación	17,45	17,72
04. Juegos y distracciones	17,76	17,63
05. La escuela	19,17	20,16
06. Habilidades docentes	12,06	13,81
07. Partes del cuerpo	27,13	26,50
08. Comidas y bebidas	26,94	26,12
Media aritmética	19,23	**19,71**

Al comparar los resultados de las dos muestras, puede advertirse que el patrón de respuestas es estrechamente similar, llegando a dibujarse una misma figura en ambos grupos. De esta manera, las líneas muestran que los promedios de palabras por cada actualizador son poco oscilantes, ya que se mantiene una continuidad mínimamente dilatada. Concretamente, los puntos referidos a los CI04, CI03 y CI07 se sobreponen casi por completo; mientras que, en los CI05, CI08 y CI01, ocurre parcialmente. En correspondencia con esto, se destacan los CI07, CI08, CI05 y CI06, cuyos puntos sobresalen en el plano, pero por razones dispares. Los tres primero asciende, representando su alta productividad léxica; contrariamente, el último desciende, por su escaso rendimiento, como se ilustra en el gráfico 33.

Gráfico 33. Comparación de los PP de los grupos de Letras H., según formato de pruebas



Con el fin de corroborar la hipótesis referida a la asociación estadística entre el Tipo de método y el caudal léxico del alumnado de Letras Hispánicas, se ha aplicado la prueba paramétrica t de Student, no sin antes determinar la homocedasticidad de los datos a través de la prueba de Levene.

Esta última arrojó que las varianzas de casi todos los grupos bajo análisis son homogéneas, salvo para el área nocional *Comidas y bebidas*, ya que esta tiene una significación de ,018 < ,05, por lo que se rechaza la H_0 y se asume las diferencias de varianzas para el CI08. Entonces, no pueden interpretarse los resultados de dicho actualizador con la prueba paramétrica.

El t-test demostró que no existe una relación estadística significativa entre la variable Tipo de método y la producción léxica de los estudiantes de Letras Hispánicas en casi todos los listados de palabras analizados, puesto que los índices de significación bilateral resultaron por encima del valor de alfa. En este sentido, se acepta la H_0, la cual admite que los datos son azarosos. Por el contrario, los cómputos correspondientes al actualizador *Habilidades y cualidades docentes* indican que sí hay una asociación entre el caudal léxico referido a dicho eje temático y la variable independiente: Tipo de método. Esto en virtud de que el valor de α = ,032 < ,05; lo que se traduce en que se rechaza la H_0. En la tabla 100 se detallan estos resultados.

Tabla 100. Resultados del t-test referido a la variable formato de prueba

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias		
		F	Sig.	t	gl	Sig. (bilateral)
Habilidades docentes	Se han asumido varianzas iguales	,016	,900	2,162	154	**,032**
	No se han asumido varianzas iguales			2,160	143,609	,032

En cuanto a los resultados sobre el CI08 obtenidos mediante la prueba no paramétrica U de Mann-Whitney, debe señalarse que *p* es igual a ,663; por lo que se constata la no existencia de asociación entre las variables dependiente e independiente.

4.3.5. Cantidad de libros leídos

En la muestra 1, correspondiente a las listas de palabras de los estudiantes de Educación Básica, quienes realizaron las pruebas en papel, se observa que el valor más alto de la media aritmética de riqueza léxica se encuentra en la variante 2 (De 6 a 10 libros) de a variable Cantidad de libros leídos, a saber: 16,75 palabras. Seguida –en orden descendente–, se encuentran las variantes 1 (De 0 a 5 libros) y 3 (Más de 10 libros), cuyos promedios son 16,46 y 16,30 lemas, correspondientemente. Así pues, entre el mayor y menor índice hay una separación de 0,45 unidades léxicas, lo que detenta poca fluctuación de los datos y, por ende, apunta a que haya pocas diferencias en el corpus. De manera global, los ejes temáticos con los PP más altos y que, por ende, sobrepasan las medias de cada variante son Partes del cuerpo, *Comidas y bebidas* y La escuela. De forma particular, al revisar los mayores □

de las áreas nocionales, se observa que en la variante 1 se encuentran los índices más altos en los actualizadores: *Comidas y bebidas* (22,54); La educación (1,95), *El profesor* (14,03) y *Habilidades y cualidades docentes* (12,92). Por su parte, la variante 2 expone los números más elevados en los CI Partes del cuerpo (24,07), La escuela (18,39), La lectura (14,32), Juegos y distracciones (14,07), como se detalla en la tabla 101.

Tabla 101. PP de Educación Básica, según *Cantidad de libros leídos*

CI	De 0 a 5	De 6 a 10	Más de 10
CI01	12,25	14,32	13,87
CI02	14,03	13,64	12,73
CI03	15,95	15,57	15,27
CI04	13,92	14,07	13,33
CI05	18,15	18,39	17,80
CI06	12,92	12,11	12,13
CI07	21,89	24,07	23,67
CI08	22,54	21,86	21,60
Media	16,46	**16,75**	16,30

En otro orden de ideas, los resultados de los encuestados de la muestra 2 –relativa a los listados de Letras Hispánicas que fueron recolectados mediante el método tradicional– muestran que el nivel de análisis 3 (Más de 10 libros) es el que expone el promedio de palabras más eminente, llegando a 19,75 lemas; es decir, 0,21 palabras por encima del valor más bajo (19,54 puntos) que lo ostenta el nivel analítico 2 (De 6 a 10 libros). Debe señalarse que, de los ocho actualizadores, los que tienen los mayores índices son Partes del cuerpo, *Comidas y bebidas* y La escuela. Según los cálculos, en la variante 1 (De 0 a 5 libros) se localizan los PP más altos de las áreas nocionales CI05, CI03, CI02 y CI06; mientras que en la variante 2, se ubica el valor más alto de CI01 (19,30 palabras); en cuanto a los restantes actualizadores –CI04, CI07 y CI08–, estos superan los cómputos en el nivel 3 (Más de 10 libros). Estos datos se detallan en la tabla 102.

Sobre la tercera muestra, los resultados presentan un orden creciente de la media de palabras desde el primer nivel de análisis (De 0 a 5 libros) –mostrando un $\overline{X}$ = 18,87 vocablos– al último nivel (Más de 10 libros), cuyo $\overline{X}$ alcanza 20,46 unidades léxicas; pasando por la variante 2 (De 6 a 10 libros), que expone un $\overline{X}$ igual a 19,11 lemas. En este sentido, se constata una distinción de 1,59 palabras entre el mayor y menor valor global de esta categoría, como se relata en la tabla 103. En correspondencia con los datos, los centros de interés Partes del cuerpo, *Comidas y bebidas*, y La escuela son los que tienen la media de palabras más altas. De manera más expedita, los CI07, CI05, CI01, CI02 y CI06 consiguen los cómputos más altos en el nivel 3; mientras que CI08 logra 27,26

palabras en el nivel 1; por último, los CI04 y CI03 tienen los PP más altos en el nivel 2. Concentrado en estos números, podría indicarse que los promedios de piezas léxicas en la muestra digital calculados para el factor Cantidad de libros leídos apuntan a una asociación sugerente entre este y el caudal léxico de los participantes de la tesis.

Tabla 102. PP de LH1, según *Cantidad de libros leídos*

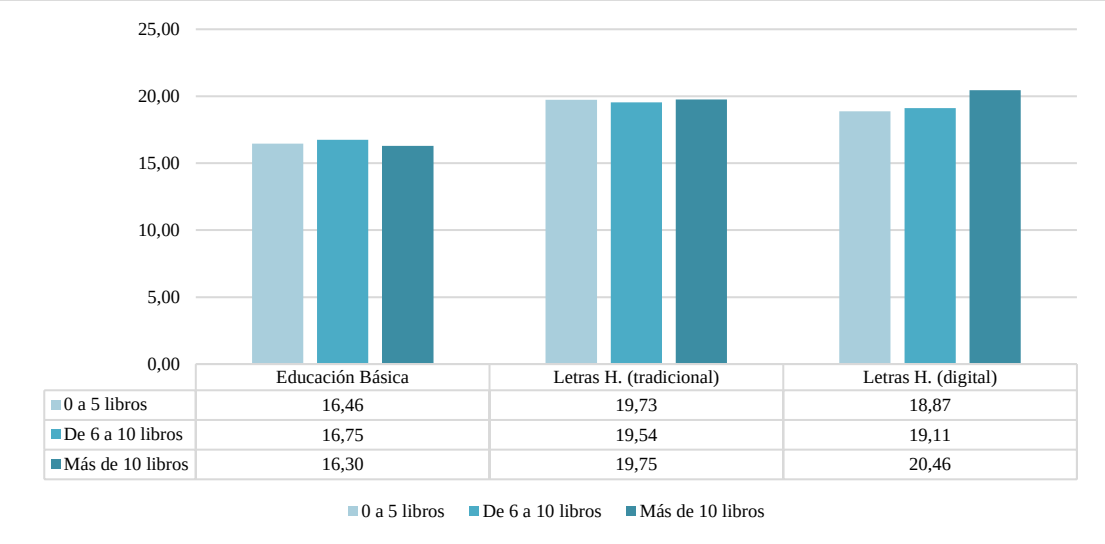
Centros de interés	De 0 a 5	De 6 a 10	Más de 10
CI01	18,54	19,30	18,47
CI02	17,39	16,20	17,17
CI03	18,43	17,80	17,03
CI04	17,11	17,50	18,17
CI05	20,86	20,50	19,40
CI06	14,39	12,90	13,57
CI07	26,00	25,70	27,23
CI08	25,11	26,40	26,97
Media	19,73	19,54	**19,75**

Tabla 103. PP de LH2, según *Cantidad de libros leídos*

Centros de interés	De 0 a 5	De 6 a 10	Más de 10
CI01	16,11	18,04	20,50
CI02	15,64	15,08	17,75
CI03	17,30	18,16	16,81
CI04	16,85	18,88	18,69
CI05	19,34	17,92	20,63
CI06	11,13	12,44	14,19
CI07	27,36	25,68	28,69
CI08	27,26	26,64	26,44
Media	18,87	19,11	**20,46**

En el gráfico 34 se trazan los resultados de las medias aritméticas generales de la variable Cantidad de libros leídos. Por una parte, este presenta la baja fluctuación de los valores de cada nivel analítico de la variable; es decir, las diferencias de los promedios de palabra no son tan marcadas de una variante a otra en las tres muestras bajo análisis. Por otra parte, más interesante aún, se observa una fuerte semejanza de los patrones de producción léxica entre las dos muestras de Letras Hispánicas, puesto que las líneas que representa a cada una están una encima de la otra, llegando a solaparse casi por completo en el nivel 2. Contrariamente, se percibe la distinción en términos de riqueza léxica entre estos dos grupos de muestras de la misma carrera y la muestra de Educación Básica, cuya línea está por debajo de las de LH.

Gráfico 34. Comparación de los PP de las tres muestras, según *Cantidad de libros leídos*



	Educación Básica	Letras H. (tradicional)	Letras H. (digital)
0 a 5 libros	16,46	19,73	18,87
De 6 a 10 libros	16,75	19,54	19,11
Más de 10 libros	16,30	19,75	20,46

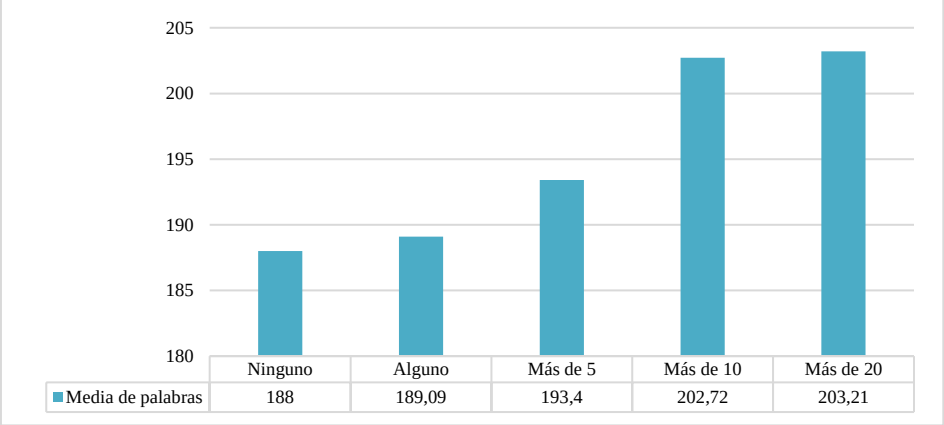
Una vez que se han descrito los contrastes de las medias aritméticas de cada muestra del estudio respecto al factor Cantidad de libros leídos, se ha procedido a determinar si este tiene alguna incidencia significativa en el caudal léxico de los participantes. Para esto se ha aplicado la prueba paramétrica Anova de un factor, no sin antes haberse calculado el grado de homocedasticidad a través del test de Levene. Este apunta a que se acepta la H_0 –lo que significa que se asume la igualdad de varianza– en todos los datos de los corpus de EB y LH1.

Una vez que se revisó la igualdad de varianzas, Anova arrojó valores de α superiores al esperado (,05) en todos los casos de los corpus de Educación Básica y Letras Hispánica (tanto en el tradicional como en el digital), tanto en el conteo global de las piezas léxicas de cada sociolecto como en los totales por áreas nocionales; de manera que se acepta la hipótesis nula para todos los subgrupos analizados en los tres corpus. De manera específica, la significación referida al número de palabras del grupo de EB es ,919. Por su parte, el valor *p* del colectivo de LH1 es ,994; mientras que la muestra digital expone un p = ,592. En este contexto, se considera que las diferencias de los datos de las tres muestras se deben al azar.

En suma, hay que resaltar el interés existente entre investigadores de distintas disciplinas, especialmente las ligadas a las ciencias pedagógicas, sociales y humanísticas, por profundizar sobre los hábitos lectores y su relación con los procesos de enseñanza-aprendizaje en los diferentes niveles de la escolaridad formal. De manera singular, la disponibilidad léxica también ha contribuido a determinar la incidencia de los consumos culturales en el vocabulario, con un propósito didáctico ulterior. Al respecto, pueden compararse muy a grandes rasgos los resultados generales del estudio de Santos Días (2017) y los de la presente tesis atinentes a la cantidad de libros leídos por los encuestados.

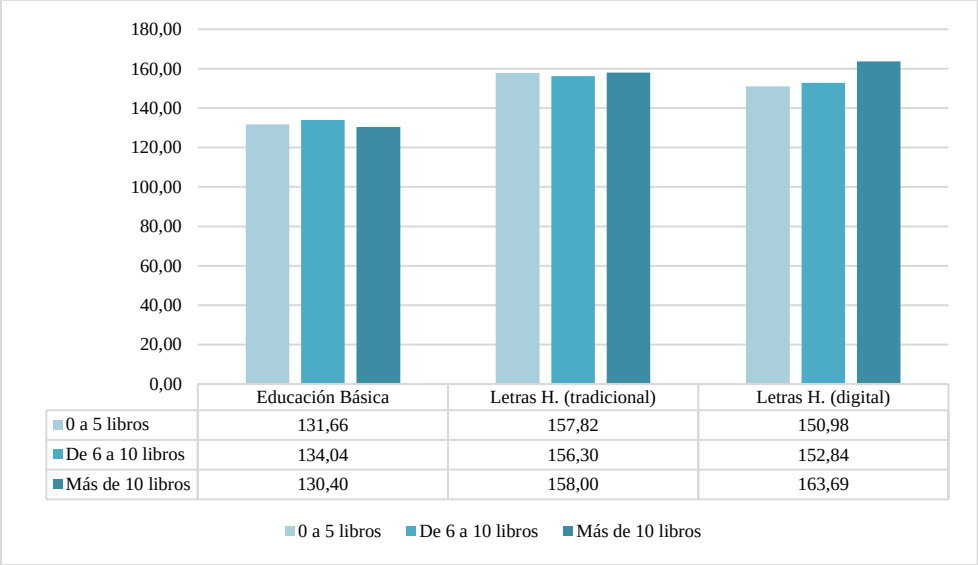
En consonancia con lo anterior, la lingüista andaluza llevó a cabo un trabajo una de cuyas finalidades era conocer la influencia de la cantidad de libros en el léxico disponible en lengua materna y extrajera de sujetos, quienes cursaban el Máster Universitario de Profesorado en la Universidad de Málaga. La variable aplicada constaba de cinco niveles cualitativos: i) Ninguno, ii) Alguno, iii) Más de 5 libros, iv) Más de 10 libros y v) Más de 20 libros. Los resultados en español relativos al promedio de palabras aportadas por los participantes indican una tendencia ascendente de productividad léxica desde el primer nivel hasta el último. Los cómputos reflejan que la media más baja se observa entre los encuestados que señalaron no haber leído ningún libro al año y se va incrementando a medida que pasa a los siguientes niveles de la variable, hasta alcanzar el valor más alto en el grupo de quienes afirmaron haber leídos más de 20 libros. Así pues, en el gráfico 38, puede apreciarse cómo va aumentando el vocabulario de los individuos a medida que aumenta la frecuencia de lectura. Guardando las diferencias habidas entre la variable de Santos Días (2017) y la de este estudio, podría decirse que, *grosso modo*, el patrón del trabajo español se replica únicamente en la muestra de Letras Hispánica (digital), según se ilustra en el gráfico 41. Puesto que se evidencia un aumento del promedio de palabras por encuestado desde la variante De 0 a 5 libros (150,98 piezas léxicas) hasta Más de 10 libros (163,69 lexías); es decir, un incremento de 12,71 lemas, lo que empuja a deducir que –al menos para este grupo– la frecuencia de lectura (medida por el número de libros leídos) incide en el caudal léxico de los hablantes.

Gráfico 35. PP según *Cantidad de libros leídos* en español del estudio de Santos Díaz (2017a)



Fuente: Elaboración propia

Gráfico 36. PP según la Cantidad de libros leídos de la presente investigación



	Educación Básica	Letras H. (tradicional)	Letras H. (digital)
0 a 5 libros	131,66	157,82	150,98
De 6 a 10 libros	134,04	156,30	152,84
Más de 10 libros	130,40	158,00	163,69

4.3.6. Frecuencia de lectura

La última variable bajo estudio es Frecuencia de lectura optativa. Esta, al igual que la previa, busca determinar el efecto de las prácticas lectoras en el léxico disponible. Específicamente, el supuesto sobre el que se basa este factor es: a mayor cantidad de horas semanales de lecturas extras al currículo de la carrera, mayor será el caudal léxico de los sujetos. En este sentido, más que el número de libros consumidos por los estudiantes, aquí se aborda el tiempo dedicado a la lectura lúdica o por placer.

En la tabla 104 pueden leerse los resultados de los promedios de palabras producidas por los estudiantes de Educación Básica en relación con la variable Frecuencia semanal de lectura opcional. En términos generales, puede observarse un aumento gradual del promedio de vocablos a medida que se avanza del nivel 1 (Ninguna hora semanal) al nivel 4 (Más de 10 horas semanales). Concretamente, se aprecian los promedios 16,23 y 18,16 lexías en las variantes de los extremos. Así pues, entre el mayor y menor índice de producción léxica se halla una diferencia de 1,93 piezas léxicas, lo que supondría un aumento del caudal léxico de los que dedican más horas a leer versus a quienes hacen lo opuesto.

No obstante, debe enfocarse el análisis en los resultados de las dos primeras variantes del factor, puesto que son las que aglutinan las respuestas del 88 % acumulado de los informantes de la muestra, como se lee en la tabla 74 de la sección 4.2.6. En este orden de ideas, se aprecia un leve ascenso de 0,22 palabras desde el nivel 1 al 2, siendo los actualizadores *Comidas y bebidas*, Partes del cuerpo y La escuela los más productivos. Además, estos tres superan la media total de cada nivel, alcanzando los promedios más altos en la segunda variable (De 1 a 5 horas). No obstante, no podría

afirmarse que hay realmente un incremento de la producción léxica en el grupo que lee más, pero sí podría suponerse una tendencia débil a la hipótesis de que, a mayor hora de lectura, mayor será el caudal léxico.

Tabla 104. PP de EB en relación con *Frecuencia de lectura opcional*

Centros de interés	Ninguno	De 1 a 5	De 6 a 10	Más de 10
01. La lectura	13,27	12,93	12,00	16,00
02. El profesor	14,33	13,56	13,11	16,75
03. La educación	15,00	15,93	15,44	16,00
04. Juegos y distracciones	12,93	13,96	13,56	16,50
05. La escuela	18,13	18,16	18,44	18,00
06. Habilidades docentes	13,40	12,29	13,22	14,75
07. Partes del cuerpo	20,80	22,74	25,22	23,50
08. Comidas y bebidas	22,00	22,06	23,78	23,75
Promedio global	129,87	131,63	134,78	145,25
Número de palabras	16,23	16,45	16,85	**18,16**

Los resultados de los alumnos de LH1 se detalla en la tabla 105. En líneas generales, –basados en los resultados de las variantes de los extremos: Ninguna hora y Más de 10 horas– podría decirse que hay un crecimiento ínfimo del caudal léxico. No obstante, el número de encuestados, quienes seleccionaron los niveles 1 y 2, es poco, apenas son el 24 % de la muestra; mientras que el 75 % de los informantes se apilan en los niveles 2 (De 1 a 5 horas) y 3 (De 6 a 10 horas). En virtud de esta distribución muestral, resulta coherente centrar la observación en los números expuestos en estas dos variantes. Así pues, se detecta que la media pasa de 20,19 palabra a 20,99; lo que podría traducirse en un ascenso mínimo (apenas 0,8 piezas léxicas). No obstante, parece más adecuado apuntar hacia una variación estable entre los dos grupos. En este corpus, son los ejes temáticos CI07, CI08, CI05 y CI01 los que reflejan los mayores índices.

Tabla 105. PP relacionada con Frecuencia de lectura de LH1

Centros de interés	**Ninguno**	**De 1 a 5**	**De 6 a 10**	**Más de 10**
01. La lectura	14,88	19,28	20,92	16,00
02. El profesor	14,25	17,44	18,08	17,00
03. La educación	15,13	18,90	18,17	14,33
04. Juegos y distracciones	13,75	18,23	18,08	17,89
05. La escuela	18,38	20,62	20,67	19,11
06. Habilidades docentes	12,38	14,05	15,67	11,56
07. Partes del cuerpo	26,13	26,28	29,42	25,11
08. Comidas y bebidas	24,25	26,69	26,92	24,11
Promedio global	139,13	161,49	167,92	145,11
Número de palabras	17,39	20,19	**20,99**	18,14

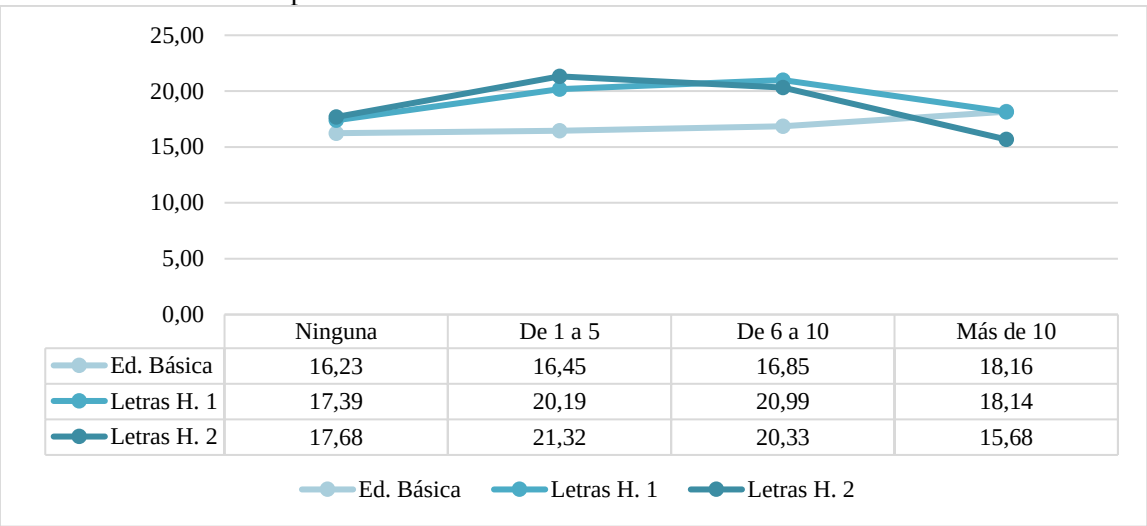
Por último, se reseñan los resultados de la media aritmética del factor Frecuencia semanal de lectura optativa en el corpus de los alumnos de Letras Hispánicas que realizaron las pruebas de disponibilidad léxica de forma remota. Al remitirse a los análisis univariantes, se detalla que el número de participantes de la variante Más de 10 horas es apenas 5, lo que representa el 5,7 % de los voluntarios de la muestra digital. En atención a lo cual, el examen se ha orientado hacia los datos de las tres primeras variantes, ya que aglutinan el 94,3 % de los informantes. Entonces, entre quienes afirmaron no leer y los que leen de 6 a 10 horas semanales, existe un aumento del caudal léxico equivalente a 2,65 palabras. Sin embargo, esta distinción se acentúa entre las variantes 1 y 2, cuya separación alcanza 3,64 palabras. Así pues, la media aritmética más alta la presenta el subgrupo que optó por la variante De 1 a 5 horas, con un $\overline{X}$ = 21,32 piezas léxicas. Justamente, en dicha variante se encuentran los mayores valores de CI07, CI08, CI04, CI01, CI03 y CI02; mientras que los actualizadores CI05 y CI06 exponen los mayores promedios en la variante De 6 a 10 horas, como se detalla en la tabla 106.

Tabla 106. PP de Frecuencia de lectura de LH2

Centros de interés	Ninguno	De 1 a 5	De 6 a 10	Más de 10
01. La lectura	14,27	20,22	17,69	13,80
02. El profesor	14,40	18,05	16,31	12,60
03. La educación	17,07	19,75	17,46	11,00
04. Juegos y distracciones	14,53	20,36	19,23	12,60
05. La escuela	19,93	20,44	21,23	15,20
06. Habilidades docentes	10,20	13,71	15,31	8,80
07. Partes del cuerpo	25,07	29,22	28,08	25,40
08. Comidas y bebidas	26,00	28,80	27,31	26,00
Promedio global	141,47	170,55	162,62	125,40
Número de palabras	17,68	**21,32**	20,33	15,68

Como complemento de estos análisis, en el gráfico 37, se contrastan los promedios de palabras generales de cada una de las muestras de la tesis. En él se detallan, en primer lugar, las semejanzas de la producción léxica entre las dos muestras de Letras Hispánicas, las cuales exponen una forma de u que contradice el supuesto de que, a más lectura, más léxico; distinción que se ve más marcada en los datos digitales. En segundo lugar, se observa la disparidad respecto a las respuestas del sociolecto de Educación Básica, que si bien presenta una producción léxica menor; sin embargo, traza una tendencia en alza.

Gráfico 37. Comparación intramuestral de los PP del factor *Frecuencia de lectura*



En virtud de que uno de los objetivos de esta tesis es determinar las posibles incidencias de las variables ligados a los consumos culturales, como la lectura, se ha realizado la prueba paramétrica de Anova de un factor a la conjunción de la variable Frecuencia de lectura optativa y la producción léxica de los encuestados. De acuerdo con los requisitos de dicho test, se estableció primero el grado de homogeneidad de las varianzas de cada una de las muestras, gracias a la operación matemática de Levene. Esta arrojó que se acepta la hipótesis nula para todos los datos de las muestras de Letras Hispánicas, tanto la del método tradicional como la digital. Es decir, para estas dos muestras, se asume la igualdad de varianzas en todos los listados de palabras analizados. Contrariamente, el test de Levene establece una desigualdad de varianza en dos de los nueve grupos de la muestra de Educación Básica, a saber: Número de palabras general (sig. ,006) y La educación (sig. ,046); por lo que no pueden interpretarse los resultados de estos dos conjuntos de datos de manera fiable con esta prueba paramétrica. No obstante, los datos de los restantes siete actualizadores de los listados de EB sí presentan el requisito de homocedasticidad.

Concretamente, para el conjunto de datos homogéneos de Educación Básica y Letras Hispánicas (tradicional), los resultados de Anova indican que no existe una relación estadística significativa entre la frecuencia de lectura optativa y el caudal léxico de los alumnos, puesto que el valor de α para todos los listados de palabras analizados es superior al esperado (,05), por lo que se acepta la H_0, y se determina el carácter azaroso de los datos.

Con el foco en la muestra levantada de manera virtual, puede señalarse que cinco de los nueve grupos analizados con Anova presentan valores de significación al nivel ,05; a saber: La lectura, con α = ,004; *Habilidades y cualidades docentes* (α = ,006); Juegos y distracciones (α = ,008); Total

general de palabras (α = ,024); y La educación (α = ,036). En estos casos, se rechaza la hipótesis nula, por lo que puede afirmarse que existe una relación estadística significativa entre la Frecuencia de lectura optativa y el léxico disponible relativo a los antes mencionados actualizadores del sociolecto de la Facultad de Letras. Estos resultados se ilustran, seguidamente, en la tabla 107.

Tabla 107. Resultados de Anova respecto a *Frecuencia de lectura* de LH2

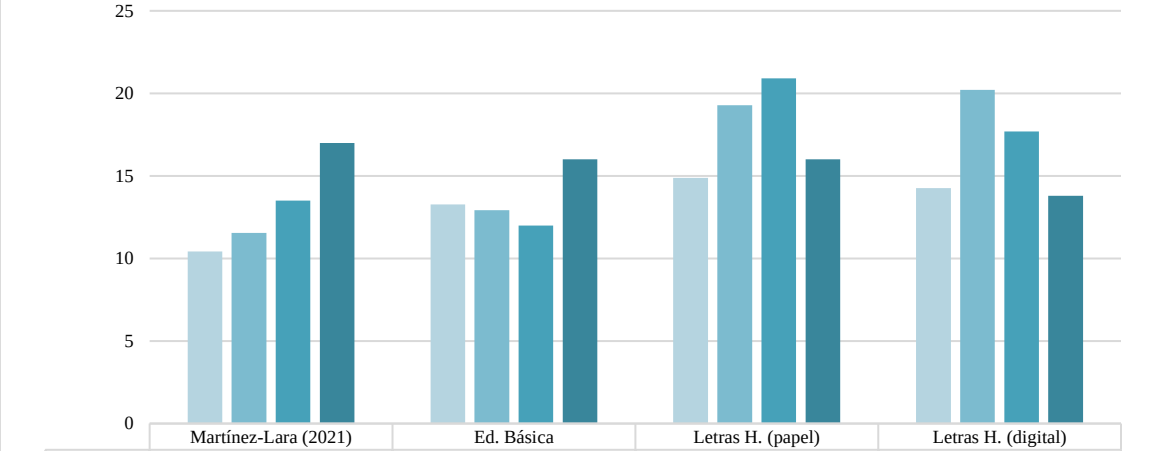
		Suma de cuadrados	gl	Media cuadrática	F	Sig.
Número de palabras	Inter-grupos	16989,335	3	5663,112	3,325	**,024**
	Intra-grupos	143047,756	84	1702,949		
	Total	160037,091	87			
La lectura	Inter-grupos	546,516	3	182,172	4,834	**,004**
	Intra-grupos	3165,303	84	37,682		
	Total	3711,818	87			
La educación	Inter-grupos	411,218	3	137,073	2,970	**,036**
	Intra-grupos	3876,600	84	46,150		
	Total	4287,818	87			
Juegos y distracciones	Inter-grupos	602,020	3	200,673	4,252	**,008**
	Intra-grupos	3963,968	84	47,190		
	Total	4565,989	87			
Habilidades docentes	Inter-grupos	301,001	3	100,334	4,498	**,006**
	Intra-grupos	1873,715	84	22,306		
	Total	2174,716	87			

En los párrafos siguientes se comparan los promedios de palabras de esta variable con los reseñados por Martínez-Lara (2021), quien también analizó la incidencia de la frecuencia de horas semanales dedicadas a la lectura optativa, en una muestra de 55 alumnos de las carreras pedagógicas de Castellano y Filosofía, Ciencias Biológicas y Parvularia de la Universidad de La Serena, Chile, sobre los ejes temáticos: *La lectura*, *La educación* y *La escuela: muebles y materiales*. Particularmente, se exponen los resultados de las medias aritméticas de palabras de los primeros ejes temáticos, ya que son analizados en las dos empresas. Además, resultaron estadísticamente significativos, según Anova, en esta tesis.

Como se aprecia en el gráfico 38, concerniente al centro de interés La lectura, los resultados de ambos estudios exponen patrones completamente dispares. En el trabajo previo –piloto del presente–, se aprecia un ascenso paulatino de los promedios de palabras a medida que se pasa de la variante 1 (Ninguna hora) a la 4 (Más de 10 horas), con una separación de 6,57 lexemas entre ambas. En las conclusiones, el investigador afirma que estos resultados, más los reportados por el Anova unifactorial, apuntan a que, ciertamente, los consumos culturales como la lectura tienen una incidencia sobre el vocabulario de los universitarios. No obstante, esta misma reflexión podría extrapolarse con cautela a los datos del corpus tradicional de Letras Hispánica, donde se reflejan un incremento de la

media de unidades léxicas desde el primer nivel al tercero (De 6 a 10 horas), con una diferencia de 6,04 lexías de distinción. Este último cómputo es semejante al del pilotaje, pero solo constituye el valor de tres de las cuatro variantes del factor.

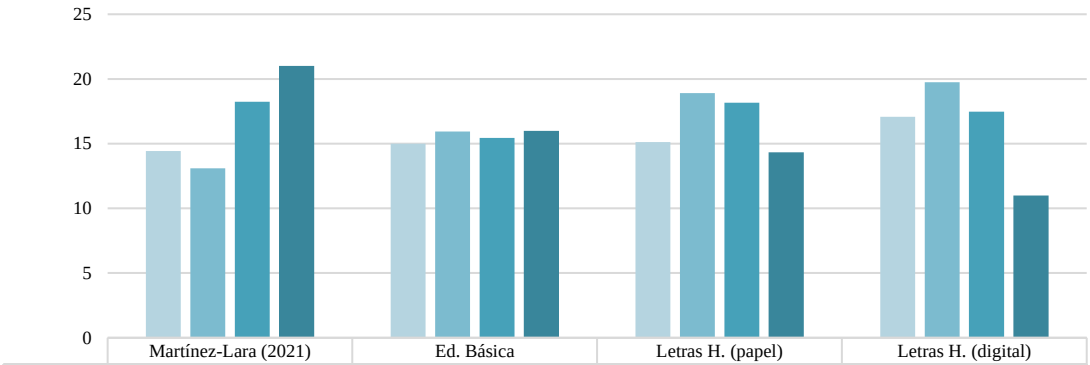
Gráfico 38. Comparación intermuestral de los PP del CI *La lectura*, según Frecuencia de lectura



	Martínez-Lara (2021)	Ed. Básica	Letras H. (papel)	Letras H. (digital)
Ninguna	10,43	13,27	14,88	14,27
De 1 a 5	11,54	12,93	19,28	20,22
De 6 a 10	13,5	12,00	20,92	17,69
Más de 10	17	16,00	16,00	13,80

Al revisar los datos comparativos del eje temático La educación, se detalla una diferencia marcada entre los estudiantes del área de Pedagogía de La Serena y los de Santiago, puesto que los primeros muestran una tendencia al alza de los promedios de palabras a medida que se sube de variante. Por el contrario, los segundos exhiben una producción léxica casi invariable, con valores que se ubican entre 15 y 16 lemas. Respecto a los grupos de Letras Hispánicas, la tendencia es inversa a los datos de los futuros pedagogos. Esto debido a que, en las muestras de educación, se marca un aumento del léxico, aunque sea mínimo, como los cómputos manifestados en la presente; mientras que, en los corpus de Letras, la propensión es a la baja, con un patrón parecido entre los listados de las pruebas digitales y tradicionales, como se contempla en el gráfico 39.

Gráfico 39. Contraste intermuestral de los PP de *La educación*, según Frecuencia de lectura



	Martínez-Lara (2021)	Ed. Básica	Letras H. (papel)	Letras H. (digital)
Ninguna	14,42	15,00	15,13	17,07
De 1 a 5	13,09	15,93	18,9	19,75
De 6 a 10	18,25	15,44	18,17	17,46
Más de 10	21	16,00	14,33	11,00

Capítulo 5. Análisis cualitativos

5.1. Consideraciones previas

Así como los estudios de disponibilidad léxica exhiben un alto componente de análisis cuantitativos –remítase al capítulo inmediatamente previo, por ejemplo–, también presentan un andamiaje cualitativo que permite observar con más detalles las características del caudal léxico más allá de los números. En este contexto, el propósito de este capítulo es, justamente, exponer los análisis cualitativos llevados a cabo en esta investigación. Antes, debe advertirse que –en lo concerniente a los estudios de DL– la exploración cualitativa de los datos es variada, puesto que han dependido en gran medida de los objetivos planteados por los lexicógrafos (Gómez-Devís, 2004). En este sentido, el cotejo de los antecedentes detalla algunos de los aspectos que se han tomado en cuenta cuando se han examinado cualitativamente los lexicones:

- Deficiencias léxicas de los encuestados (Paredes, 1999; Trigo Ibáñez *et al.*, 2018; Nalesso, 2022);
- transferencias léxicas (Álvarez y López, 2021; Gómez Devís y Serrano, 2021; de la Maya y López, 2021);
- dialectalismos (Valencia, 2011; Trigo Ibáñez, 2011);
- léxico especializado (Santos Díaz, 2020; Marcos-Calvo y Herranz, 2021; Castillo y Santos Díaz, 2021);
- redes asociativas y prototipos (Echeverría *et al.*, 2008; Ferreira y Echeverría, 2010; Ávila Muñoz y Sánchez Sáez, 2014; Gómez Molina, 2021; Blanco *et al.*, 2020), entre otros rasgos.

Si bien, considerar estos enfoques en conjunto sería ideal en una investigación; la realidad apunta a que resultaría casi inabarcable. Así pues, en esta tesis, se optó únicamente por determinar las convergencias y divergencias de los caudales léxicos de los encuestados, a tenor de las variables del estudio: *sexo*, *carrera*, *año de curso*, *formato de prueba* y *frecuencia de lectura optativa*. En esta línea, los análisis han permitido conocer y describir las piezas léxicas que caracterizan a un grupo respecto del otro. Paralelamente, se ha intentado explicar las vinculaciones entre el vocabulario y los rasgos extralingüísticos de los participantes.

Según el arqueo bibliográfico, no ha habido unanimidad en el rango de corte de las palabras más disponibles; al contrario, han dependido de los objetivos de cada investigación. Así, en algunas pesquisas se han analizado las cinco, siete o diez piezas léxicas más disponibles, como Fregoso-Peralta y Aguilar-González (2022), Pacheco Mirabal *et al.* (2016) y Pacheco *et al.* (2017), respectivamente.

También, se han contrastados las primeras cincuenta, como en los trabajos de Gómez Devís (2004), Martínez (2007), Trigo Ibáñez y González (2010) y Urzúa (2018). Sobre el español de Chile, Valencia (1998-1999; 2010), Valencia y Echeverría (1999) han analizado los cien primeros vocablos con los mayores IDL. En otras investigaciones, el corte se ha basado en valores propuestos a partir del índice de disponibilidad léxica, el porcentaje de aparición y la frecuencia acumulada (Santos Díaz, 2017b). Con relación a esta forma, pueden mencionarse los estudios de Pérez (2015) y Zhou (2021), quienes tomaron en cuenta los vocablos con IDL ≥ 0,05; mientras que Hidalgo (2017) y Martínez-Lara (2021) realizaron el corte a partir de las lexías con IDL ≥ 0,1. No obstante, en este trabajo, se compararon las veinte primeras palabras de cada CI, en consonancia con algunos antecedentes, tales como Prado y Galloso (2015), Santos Díaz (2015) y Herranz (2020).

A manera de recapitulación, los datos léxicos provienen de tres muestras: las dos primeras basadas en los lexicones de estudiantes de Educación Básica (EB) y Letras Hispánicas (LH1), quienes contestaron las pruebas de DL en papel. Mientras que la tercera muestra está integrada por los listados de palabras de alumnos de Letras Hispánicas (LH2), quienes respondieron los cuestionarios de forma digital. Entonces, en los epígrafes siguientes, se exponen las comparaciones de las veinte primeras palabras de los centros de interés que resultaron estadísticamente significativos, según las variables *sexo*, *carrera*, *año de curso*, *formato de prueba* y *frecuencia de lectura optativa*.

Asimismo, hay que acotar que se exploraron, cualitativamente, los vocabularios de los sociolectos que resultaron estadísticamente significativos, según las pruebas paramétricas (t de Student, Anova) y no paramétricas. En las tablas comparativas, se muestran los contrastes a tenor de los siguientes atributos tipográficos: i) se hallan en letra redonda las lexías compartidas entre los grupos, y ii) en **negrita** las palabras distintivas, como se detallan a continuación en el subapartado 5.2.

5.2. Convergencias y divergencias del léxico disponible

5.2.1. Variable sexo

En el caso de la variable *sexo*, solo dos de los ocho centros de interés resultaron estadísticamente significativos, conforme al t-test; estos son: *La escuela: muebles y materiales*, de los listados de Educación Básica; y, *Partes del cuerpo*, de LH1.

La escuela: muebles y materiales

Los veinte vocablos más disponibles, según el sexo de los informantes de Educación Básica, del actualizador *La escuela: muebles y materiales*, se ilustran en Tabla 108.

Tabla 108. Los veinte vocablos más disponibles de *La escuela*, de EB

	Hombres		**Mujeres**	
	Vocablo	IDL	Vocablo	IDL
1	mesa	0,6744	mesa	0,6842
2	silla	0,6439	silla	0,6590
3	lápiz	0,5006	lápiz	0,6228
4	pizarra	0,4689	pizarra	0,5069
5	plumón	0,4286	cuaderno	0,4637
6	libro	0,3167	libro	0,3722
7	cuaderno	0,3055	plumón	0,3023
8	escritorio	0,2731	estuche	0,2765
9	borrador	0,2307	goma (de borrar)	0,2444
10	estuche	0,2075	escritorio	0,2290
11	goma (de borrar)	0,1861	**regla**	0,2058
12	**ventana**	0,1805	estante	0,1691
13	compu(tador/a)	0,1791	**mochila**	0,1554
14	estante	0,1737	borrador	0,1525
15	**pupitre**	0,1663	tijera	0,1503
16	**pizarrón**	0,1505	**destacador**	0,1484
17	tijera	0,1375	corrector	0,1373
18	**hoja**	0,1340	**cartulina**	0,1238
19	corrector	0,1278	compu(tador/a)	0,1193
20	**proyector**	0,1274	**pegamento**	0,1134

En la tabla 108, se aprecia que, de manera creciente, las piezas léxicas presentan IDL > 0,1, siendo la palabra *proyector* el que ocupa el rango 20 en el listado de los hombres, con IDL = 0,1274; mientras que en el de las mujeres, en ese mismo rango, se halla *pegamento*, con IDL = 0,1134. Por su parte, el vocablo *mesa* es el que corona los lexicones tanto del sociolecto masculino como del femenino, con IDL equivalentes a 0,6744 y 0,6842, respectivamente. Entonces, dicho lema es el núcleo del actualizador. Debe resaltarse que, en los listados de ambos sexos, se observan las mismas palabras, con el mismo rango de aparición, en los primeros cuatro lugares; estas son, en orden descendente: *mesa*, *silla*, *lápiz* y *pizarra*. Esto desvela las altas similitudes entre ambos grupos. En relación con este punto, ambos listados encajan en ¾ partes de las palabras bajo análisis, ya que quince de las veinte palabras son idénticas en los dos sexos. Opuestamente, cada sociolecto muestra cinco voces disímiles, la de los hombres son: *ventana* (puesto 12), *pupitre* (15), *pizarrón* (16), *hoja* (18) y *proyector* (20). En cambio, las de las mujeres son: *regla* (11), *mochila* (13), *destacador* (16), *cartulina* (18) y *pegamento* (20). A grandes rasgos, puede apreciarse que existe una diferencia sobre el tipo de palabras diferentes evocadas por cada grupo, ya que, mientras los varones escribieron lemas relacionados con

estructuras de construcción, muebles pesados y materiales básicos. En contraste, las chicas produjeron lexías afines a materiales de uso práctico y sencillo y del contenedor de ellos.

Partes del cuerpo

Primeramente, debe reseñarse que los veinte primeros vocablos más disponibles de los estudiantes de Letras Hispánicas empiezan –de forma ascendente– con IDL > 0,1, llegando a alcanzar un IDL > 0,6; mostrando el valor más alto en el sociolecto masculino (*ojo* IDL = 0,6286), en comparación con el femenino (*mano* IDL = 0,6042). Igualmente, puede observarse que las piezas léxicas (*ojo* y *mano*) recién mencionadas representan los elementos nucleares de cada grupo, ya que encabezan los listados con los índices más altos. En el caso de las féminas, el primer lugar lo ocupa *mano*, seguida por *ojo*, *brazo* y *cabeza*; mientras que los lexicones de los varones está a la cabeza *ojo*, seguido por *mano*, *cabeza* y *pie*. Sobre el grado de coincidencia, puede señalarse que este es alto, puesto que ambos sexos comparten dieciséis de los veinte vocablos bajo análisis; mientras que discrepan en tan solo cuatro lexemas. Por parte de las mujeres, se encuentran *rodilla* (en el rango 13), *espalda* (17), *codo* (18) y *estómago* (19); es decir, partes externas de la fisionomía humana y un órgano interno. Por su parte, las cuatro lexías peculiares de los hombres son *pene* (11), *pulmón* (15), *hígado* (17) y *lengua* (18); las cuales remiten a órganos, tres internos y uno externos. En la tabla 102, a continuación, se detalla este contraste.

Tabla 109. Los 20 vocablos más disponibles *Partes del cuerpo*, de LH1

	Mujeres		**Hombres**	
	Vocablo	IDL	Vocablo	IDL
1	mano	0,6042	ojo	0,6286
2	ojo	0,5825	mano	0,5502
3	brazo	0,5373	cabeza	0,4432
4	cabeza	0,5338	pie	0,4286
5	pierna	0,5238	pierna	0,4272
6	pie	0,4784	pelo	0,3442
7	dedo	0,4432	brazo	0,3319
8	uña	0,3688	nariz	0,3306
9	nariz	0,3650	dedo	0,2844
10	boca	0,3606	oreja	0,2802
11	oreja	0,3481	**pene**	0,2330
12	pelo	0,2929	corazón	0,2287
13	**rodilla**	0,2576	uña	0,2215
14	corazón	0,2522	cerebro	0,2179
15	cuello	0,1891	**pulmón**	0,2030
16	diente	0,1503	boca	0,1940

17	**espalda**	0,1502	**hígado**	0,1702
18	**codo**	0,1462	**lengua**	0,1575
19	**estómago**	0,1417	diente	0,1463
20	cerebro	0,1415	cuello	0,1441

En este orden de ideas, hay que destacar que los veinte primeros vocablos más disponibles, tanto de las mujeres como de los hombres, son sustantivos que se ligan directamente con la fisionomía. En otras palabras, las voces evocadas están inequívocamente relacionadas con el campo nocional planteado; no han aparecido en este primer corte unidades léxicas diferentes a la anatomía humana. Pero sí es evidente que los participantes trajeron a la memoria desde órganos internos (*corazón*, *estómago*, *pulmón*) y externos (*nariz*, *oreja*, *boca*, *pene*); extremidades (*mano*, *brazo*, *pierna*, *pie*); células de distintos tipos (*uña*, *pelo*) hasta huesos (*diente*).

5.2.2. Variable carrera

En esta tesis se han descrito los lexicones de alumnos de dos carreras: Educación Básica y Letras Hispánicas, puesto que uno de los objetivos es comparar el vocabulario de cada uno de estos grupos. Para esto, únicamente se han tomado en cuenta los datos recogidos mediante el método tradicional o en papel; es decir, las muestras 1 y 2 que integran el subcorpus 1. Con base en los resultados del t de Student, los actualizadores estadísticamente significativos son: *la lectura*, *La escuela: muebles y materiales* y *Comidas y bebidas*. Por su parte, *el profesor*, *juegos y distracciones* y *partes del cuerpo*, resultaron significativo a través de la prueba no paramétrica U de Mann-Whitney.

La lectura

El contraste entre las veinte palabras más disponibles concernientes al CI *la lectura* muestra que, en primer lugar, *libro* es la que exhibe los índices de disponibilidad léxica más altos en las dos muestras, siendo mayor en los datos de Letras Hispánicas en relación con los de Educación Básica, 0,8779 y 0,7500, respectivamente. Debido a lo anterior, se ubica en la cabeza de las listas. En este sentido, podría afirmarse que este lexema corresponde al núcleo de este campo nocional. A este le siguen de forma descendente los lemas *letra* y *leer*, los cuales difieren en los puestos (2^{do} y 3^{ero}) de cada conjunto. En alusión a los IDL, los últimos elementos léxicos de este corte tienen valores por encima de 0,07; muy similares en ambos sociolectos; concretamente, *personaje* (IDL = 0,0755, en EB) y *paper* (IDL = 0,0758, en LH). Sobre LH, llama mucho la atención que de la posición veinte a la diecinueve hay un salto del índice equivalente a 0,0172, puesto que el vocablo *escritura* (19) tiene un IDL = 0,0930; al contrario, en EB el índice va subiendo de forma gradual, llegando a IDL > 0,09

en la posición 15. Asimismo, la muestra de pedagogía 11 vocablos con IDL > 0,01; mientras que los humanistas contabilizan 17, como puede apreciarse en la tabla 110.

Tabla 110. Las 20 palabras más disponibles del CI *La escuela*, según Carrera

	Ed. Básica		**Letras**	
	Vocablo	**IDL**	**Vocablo**	**IDL**
1	libro	0,7500	libro	0,8779
2	leer	0,2875	letra	0,3569
3	letra	0,2759	leer	0,2667
4	palabra	0,2122	autor	0,2023
5	cuento	0,2007	palabra	0,1977
6	texto	0,1688	biblioteca	0,1809
7	autor	0,1447	página	0,1680
8	historia	0,1330	**lápiz**	0,1627
9	**comprensión**	0,1208	cuento	0,1529
10	novela	0,1141	lector	0,1329
11	biblioteca	0,1026	novela	0,1107
12	**conocimiento**	0,0964	texto	0,1088
13	página	0,0957	**literatura**	0,1077
14	**revista**	0,0921	**lentes**	0,1035
15	**lenguaje**	0,0906	historia	0,1034
16	**noticia**	0,0838	**poesía**	0,1016
17	**párrafo**	0,0802	**escritor**	0,1011
18	**tiempo**	0,0772	**hoja**	0,0936
19	lector	0,0772	**escritura**	0,0930
20	**personaje**	0,0755	**paper**	0,0758

Al analizar los cinco primeros lexemas y, por ende, los más disponible, se observa que (además de los anteriormente mencionados) se cuentan los vocablos *palabra*, *cuento* y *autor*. Los dos últimos difieren de un colectivo a otro, en este corte. Esta distinción podría señalar que –para los discentes de pedagogía– es más disponible una unidad de cita referida a géneros literarios, mientras que –para los de letras– pareciera que tiene mayor potencialidad una lexía relacionada con los elementos materiales del proceso comunicativo escrito (Escandell, 2019).

En segundo lugar, doce de las veinte entradas léxicas comparadas coinciden en los dos sociolectos. Por lo tanto, las divergentes suman ocho. Estas son –en los datos de EB– *comprensión lectora*, la cual apareció en el rango 9; prosiguen *conocimiento* (12), *revista* (14), *lenguaje* (15), *noticia* (16), *párrafo* (17), *tiempo* (18) y *personaje* (20). Por su parte, la lista de LH presenta las siguientes piezas léxicas únicas: *lápiz*, ubicada en el rango 8; continúan *literatura* (13), *lentes* (14), *poesía* (16), *escritor* (17), *hoja* (18), *escritura* (19) y *paper* (20). A grandes rasgos, podría

indicarse que pareciera que las unidades de cita del grupo de EB se inclinan más hacia aspectos cognitivos de la lectura; géneros y estructuras textuales (Parodi, 2004; Venegas *et al.*., 2016; Ibáñez *et al.*, 2017); y, elementos pragmáticos materiales e inmateriales (Escandell, 2019). En el caso de LH, los vocablos distintivos sugerirían una vinculación con útiles y elementos materiales de la comunicación; además de procesos cognitivos y géneros textuales.

El profesor

En el centro de interés *El profesor*, de los veinte vocablos más disponibles de los dos grupos bajo análisis, pueden identificarse doce convergentes. Estos lexemas son: *enseñar*, *vocación*, *conocimiento*, *colegio*, *enseñanza*, *docente*, *educación*, *aprendizaje*, *alumno*, *aprender*, *autoridad* y *clase*, con IDL > 0,07. Sin embargo, su distribución en las posiciones de cada listado es distinta. En el conjunto de Educación Básica, la palabra con el IDL más elevado es *enseñar*; seguida por *vocación* y *guía*; mientras que, en el grupo de Letras Hispánica, la unidad léxica con el mayor IDL es *clase*; la cual es seguida por *universidad* y *colegio*. Es decir, considerando las lexías de estos tres lugares de cada listado, se observa una clara diferencia entre el léxico con el que cada grupo identifica al referente de esta área nocional. En la tabla 111 se ilustran estas palabras.

Tabla 111. Las 20 palabras más disponibles del CI El profesor, según Carrera

	Ed. Básica		Letras	
	Vocablo	IDL	Vocablo	IDL
1	enseñar	0,2293	clase	0,2569
2	vocación	0,2118	**universidad**	0,2556
3	**guía**	0,1637	colegio	0,2353
4	conocimiento	0,1499	enseñanza	0,2163
5	**educador**	0,1459	conocimiento	0,2159
6	**paciencia**	0,1321	**pizarra**	0,2091
7	colegio	0,1319	alumno	0,1854
8	enseñanza	0,1190	**prueba**	0,1769
9	docente	0,1088	enseñar	0,1731
10	educación	0,1047	**estudio**	0,1490
11	aprendizaje	0,0937	vocación	0,1448
12	alumno	0,0934	**estudiante**	0,1283
13	aprender	0,0930	autoridad	0,1256
14	**apoyo**	0,0925	aprendizaje	0,1239
15	autoridad	0,0909	aprender	0,1208
16	**buena onda**	0,0903	docente	0,1165
17	clase	0,0870	educación	0,1155
18	**maestro**	0,0858	**plumón**	0,1026
19	**ayuda**	0,0843	**sala de clase**	0,0989

20	**persona**	0,0790	**nota**	0,0869

Sobre las palabras divergentes en cada grupo, los estudiantes de EB reportaron las siguiente ocho: *guía*, *educador*, *paciencia*, *apoyo*, *buena onda*, *maestro*, *ayuda* y *persona*. Por su parte, los alumnos de LH escribieron: *universidad*, *pizarra*, *prueba*, *estudio*, *estudiante*, *plumón*, *sala de clase* y *nota*. A partir de una observación superficial, puede señalarse que estás distinciones exponen claramente los puntos de vistas contrarios que cada grupo de estudiante tiene sobre el referente profesor. Puesto que, los miembros de la Facultad de Educación resaltan los roles y valores del personal académico, mientras que los de la Facultad de Letras subrayan los instrumentos, lugares y medios utilizados por los docentes para cumplir su labor.

En cuanto a las categorías gramaticales, los estudiantes de ED escribieron 17 sustantivos (*vocación*, *conocimiento*, *colegio*, *enseñanza*, *docente*, *educación*, *aprendizaje*, *alumno*, *autoridad*, *clase*, *guía*, *educador*, *paciencia*, *apoyo*, *maestro*, *ayuda* y *persona*); 2 verbos (*enseñar* y *aprender*), y 1 adjetivo (*buena onda*). Por su parte, los alumnos de LH enumeraron 18 sustantivos (*vocación*, *conocimiento*, *colegio*, *enseñanza*, *docente*, *educación*, *aprendizaje*, *alumno*, *autoridad*, *clase*, *universidad*, *pizarra*, *prueba*, *estudio*, *estudiante*, *plumón*, *sala de clase* y *nota*) y 2 verbos (*enseñar* y *aprender*).

Juegos y distracciones

En el mismo contexto de la variable *carrera*, el análisis estadístico arrojó como significativa la relación de los listados de palabras del actualizador *juegos y distracciones*, por lo que se presenta el análisis contrastivo de los veinte vocablos más disponibles de cada sociolecto. En la tabla 112, a continuación, se detallan estas unidades léxicas.

Tabla 112. Las 20 palabras más disponibles del *Juegos y distracciones*, según *Carrera*

	EB		**Letras**	
	Vocablo	**IDL**	**Vocablo**	**IDL**
1	diversión	0,2004	compu(tador/a)	0,2757
2	**entretención**	0,1687	carta	0,2052
3	juego de mesa	0,1602	leer	0,1990
4	celular	0,1521	celular	0,1942
5	amigo	0,1431	película	0,1854
6	**Monopoly**	0,1342	ocio	0,1839
7	**niño**	0,1267	**música**	0,1796
8	película	0,1080	videojuego	0,1789
9	tiempo libre	0,1009	tiempo libre	0,1598
10	compu(tador/a)	0,1008	**tele(visión)**	0,1349

11	videojuego	0,1006	serie	0,1332
12	carta	0,0945	diversión	0,1300
13	**escondida**	0,0850	**internet**	0,1297
14	ocio	0,0842	**libro**	0,1281
15	**Play(Station)**	0,0812	amigo	0,1141
16	leer	0,0801	**dormir**	0,1114
17	pintar	0,0799	juego de mesa	0,1104
18	serie	0,0790	**lectura**	0,0921
19	**fútbol**	0,0758	**LOL**	0,0807
20	**pelota**	0,0744	pintar	0,0802

Resulta interesante que las últimas piezas léxicas de este conjunto tengan IDL mayores a 0,07; exactamente, *pelota* (IDL = 0,0744), en el grupo de EB, y *pintar* (IDL = 0,0802), en el de LH. Estos datos muestran a simple vista que el grupo de humanidades supera a los de pedagogía en este índice. En cuanto a las palabras más disponibles –y que, por ende, se ubican a la cabeza de los listados–, estas son: *diversión* (IDL = 0,0744), en el de EB, y *compu(tador/a)*, en el de LH. Por lo que, a grandes rasgos, puede decirse que estas pueden considerarse como los elementos nucleares de cada sociolecto, en consideración al área nocional *juegos y distracciones*.

Si se enfoca el análisis únicamente en las cinco primeras unidades léxicas, se puede observar una diferencia entre el tipo de palabras citadas por cada colectivo. Esto debido a que, por una parte, los alumnos de EB escribieron: *diversión*, *entretención*, *juego de mesa*, *celular* y *amigo*; las cuales parecieran abarcar posibilidades de pasatiempos más amplias y generales. Adicionalmente, evocan –podría señalarse– sinónimos del nombre del CI; hiperónimos, dispositivos digitales y entes con los que suele relacionarse el ocio. Por otra parte, los discentes de LH reportaron: *compu(tador/a)*, *carta*, *leer*, *celular* y *película*. Es decir, palabras que sugieren formas más concretas de distensión, ya que apuntan un poco más directamente a elementos y procesos estimados como bureo, a saber: artefactos tecnológicos, materiales y procesos cognitivos. Asimismo, comparativamente, el vocablo *celular* coincide en ambos grupos, ocupando el mismo rango (4). Paralelamente, de este pequeño corte de las listas, puede señalarse que –en relación con las unidades léxicas de EB– las de LH incluyen no solo sustantivos, sino también un verbo, que se vincula directamente con las actividades propias de los universitarios y, aún más, con las de los alumnos de Letras Hispánicas.

En este mismo orden de ideas, ambos sociolectos coinciden en trece de las veinte palabras más disponibles. El primer lema distintivo que aparece en la lista de EB es *entretención* (rango 2), seguido por *Monopoly* (6), *niño* (7), *escondida* (13), *Play(Station)* (15), *fútbol* (19) y *pelota* (20). Así pues, estas –aparte del sinónimo de la denominación del actualizador– agrupan divertimentos, tanto de mesa

como al aire libre o espacios amplios; dispositivos y entes emparejados con el campo lúdico. Por su parte, los alumnos de LH exhiben como léxico contrastivo: *música*, *tele(visión)*, *internet*, *libro*, *dormir*, *lectura* y *LOL*. Estas podrían considerarse como evocación de formas de entretenimiento enfocadas en actividades, personales e individuales, de desarrollo intelectual, estratégico y sedentario. De alguna manera, podría afirmarse que estas piezas léxicas nucleares reflejan aparentemente más dinamismo en el caso de EB; mientras que, opuestamente, las de LH parecieran dar la impresión de pasividad o sedentarismo.

La escuela: muebles y materiales

Primero, debe resaltarse que, tanto en los datos de Educación Básica como en los de Letras Hispánicas, *mesa* y *silla* coinciden –en este orden– en el primer y segundo puesto de los listados de ambos sociolectos. Por lo tanto, estas unidades de citas pueden considerarse nucleares en el eje temático *La escuela: muebles y materiales*, siendo *mesa* la que presenta los índices de disponibilidad léxica más altos en los dos grupos, 0,6497 en EB; y 0,7406 en LH. En este orden de ideas, los lemas del grupo de EB inician –de forma ascendente– con IDL por encima de 0,09, específicamente: *casillero*, rango 20, con IDL = 0,0969; este valor va incrementándose paulatinamente a medida que se va subiendo hasta el rango 1. Por su parte, el listado de LH inicia el conteo ascendente con un valor de IDL mayor a 0,1; concretamente, el vocablo *corrector*, rango 20, cuyo IDL = 0,1358. Es decir, las veinte palabras más disponibles de LH tienen IDL > 0,1; mientras que EB suma diecinueve, como se aprecia en la tabla 113.

Tabla 113. Las 20 palabras más disponibles del CI *La escuela*, según *Carrera*

	Educación básica		**Letras Hispánicas**	
	Vocablo	**IDL**	**Vocablo**	**IDL**
1	mesa	0,6497	mesa	0,7406
2	silla	0,6225	silla	0,7108
3	pizarra	0,5298	lápiz	0,6244
4	lápiz	0,3864	pizarra	0,5776
5	plumón	0,3598	cuaderno	0,4717
6	cuaderno	0,2549	plumón	0,4077
7	estuche	0,2144	estuche	0,3545
8	libro	0,1993	goma (de borrar)	0,3308
9	escritorio	0,1882	libro	0,3002
10	sala de clase	0,1859	borrador	0,2373
11	**cartulina**	0,1848	proyector	0,2070
12	**ventana**	0,1807	**regla**	0,1912
13	goma (de borrar)	0,1601	**mochila**	0,1835

14	**patio**	0,1472	compu(tador/a)	0,1802
15	compu(tador/a)	0,1284	escritorio	0,1794
16	proyector	0,1257	**pegamento**	0,1670
17	borrador	0,1216	tijeras	0,1651
18	**estante**	0,1207	sala de clases	0,1486
19	tijera	0,1065	**destacador**	0,1394
20	**casillero**	0,0969	**corrector**	0,1358

Y segundo, al analizar las cinco primeras piezas léxicas más disponibles, se observa que en ambos sociolectos se hallan –además de *mesa* y *silla*– *pizarra* y *lápiz*; pero no se localizan los vocablos *plumón* (EB) y *cuaderno* (LH). Sin embargo, estos últimos sí aparecen entre los veinte. De hecho, los datos de las dos carreras son afines en quince de los veinte lexemas de la lista, distinguiéndose únicamente por cinco. Así pues, en el conjunto de EB, se identifican los siguientes: *cartulina*, posición 11; *ventana* (12), *patio* (14), *estante* (18) y *casillero* (20). En relación con LH, las unidades léxicas diferentes son *regla* (12), *mochila* (13), *pegamento* (16), *destacador* (19) y *corrector* (20). Con relación a esta muestra de palabras peculiares podría indicarse que se aprecia una clara particularidad entre las dos carreras, puesto que los encuestados de pedagogía aportaron vocablos que tocan varios aspectos del mundo escolar, no solo los concernientes al mobiliario y los utensilios, sino también a la infraestructura y espacios de los colegios. En tanto que los participantes de LH proporcionaron lexías correspondientes a materiales solamente. En atención a lo anterior, podría hipotetizarse que este pequeño corte podría reflejar las distintas concepciones que se tienen acerca de la esfera del colegio en relación con las áreas de especialidad.

Partes del cuerpo

Con base en los promedios de palabra, *Partes del cuerpo* es el actualizador más productivo de esta tesis. Paralelamente, es el más compacto, según los índices de cohesión y densidad léxica. En el conteo de los veinte lemas más disponibles, los vocablos *mano* y *ojo* resultaron con los valores más altos. En el listado de Educación Básica, los dos aparecen en esa secuencia; es decir, *mano*, con un IDL = 0,5983, ocupa el rango 1. Opuestamente, *ojo*, con un IDL = 0,7183, es el que ostenta el mayor cómputo en el grupo de Letras Hispánicas. Entonces, ambas unidades léxicas pueden considerarse como las centrales en cada sociolecto. En la tabla 114, se detallan los resultados.

Tabla 114. Las 20 palabras más disponibles del CI *Partes del cuerpo*

	Educación Básica		Letras Hispánicas	
	Vocablo	IDL	Vocablo	IDL
1	mano	0,5983	ojo	0,7183

2	ojo	0,5958	mano	0,6730
3	cabeza	0,5208	nariz	0,6076
4	pierna	0,5139	pie	0,5466
5	brazo	0,5072	cabeza	0,5306
6	pie	0,4772	brazo	0,5288
7	dedo	0,4227	dedo	0,5235
8	nariz	0,3649	pierna	0,5073
9	uña	0,3485	uña	0,4826
10	oreja	0,3418	oreja	0,4691
11	boca	0,3366	boca	0,4384
12	pelo	0,3035	pelo	0,3735
13	corazón	0,2516	cuello	0,3652
14	rodilla	0,2318	rodilla	0,3498
15	cuello	0,1838	**hombro**	0,3032
16	cerebro	0,1579	diente	0,2865
17	diente	0,1530	corazón	0,2705
18	**espalda**	0,1472	**lengua**	0,2654
19	**codo**	0,1465	**pecho**	0,2595
20	**pene**	0,1438	cerebro	0,2479

En el corte de las cinco primeras palabras, se aprecia que los dos conjuntos de datos coinciden en los lexemas *mano*, *ojo* y *cabeza*; pero son dispare respecto a *pierna* y *brazo* –que se hallan en EB–, y *nariz* y *pie*, en LH. Así pues, podría suponerse que los pedagogos potenciaron las extremidades, mientras que los letrados evocaron zonas variadas de la fisionomía humana. Sin embargo, se hace la acotación de que estos vocablos distintivos sí se localizan entre los veinte más disponibles de cada sociolecto. Al continuar con la exploración, se detalla que las unidades léxicas de este actualizador son altamente coincidentes, debido a que comparten diecisiete y, únicamente, se oponen en tres, a saber: *espalda*, *codo* y *pene*, ubicadas en este orden descendentemente en los últimos puestos del listado de EB. Por su parte, los lexemas exclusivos de LH son *hombro*, rango 15, *lengua* (18) y *pecho* (19). En consonancia con estos datos, podría decirse que ambos grupos –además de tener una alta convergencia léxica en este eje temático– aportan palabras directamente ligadas al cuerpo.

Por último, es importante apuntar que los valores de los IDL de los vocablos del CI07 son bastante altos, en relación con los manifestados en los demás CI que resultaron estadísticamente significativos. En este marco, hay que reseñar que las últimas piezas léxicas de cada carrera exhiben IDL > 0,14; valor que va aumentando a medida que los vocablos pasan a los rangos superiores. De hecho, el lexema *pene*, localizado en el puesto 20 de Educación Básica, presenta un IDL = 0,1438; mientras que el 20mo elemento léxico de Letras Hispánicas, *cerebro*, tiene un IDL = 0,2479.

Comidas y bebidas

Por último, se analizan las veinte unidades léxicas más disponibles del área nocional *Comidas y bebidas*, según la variable *carrera*. Primeramente, debe acotarse que los dos sociolectos –estudiantes de Educación Básica y Letras Hispánicas– coinciden en que la lexía *Coca-Cola* se encuentra a la cabeza de sus respectivos listados, ya que resultó con los valores más altos de disponibilidad. Concretamente, el primer grupo alcanzó un IDL = 0,4829; mientras que el segundo muestra un IDL = 0,4645. En esta ocasión, ambas cifras son bastante similares, siendo mayor entre los pedagogos. En este sentido, puede apuntarse que *Coca-Cola* es el vocablo nuclear de este actualizador, en conformidad con las respuestas de los dos sociolectos bajo análisis. Al considerar el índice de DL, se advierte que los valores del vocablo localizado en el rango 20 de cada carrera son mayores a 0,1. Para más detalle, véase la tabla 115.

Tabla 115. Las 20 palabras más disponibles del CI *Bebidas y comidas*

	Educación Básica		Letras Hispánicas	
	Vocablo	IDL	Vocablo	IDL
1	Coca-Cola	0,4829	Coca-Cola	0,4645
2	arroz	0,3331	agua	0,4274
3	fideo	0,3010	arroz	0,3754
4	papa frita	0,2970	jugo	0,3684
5	pizza	0,2917	carne	0,3019
6	jugo	0,2823	fideo	0,2898
7	agua	0,2589	pizza	0,2760
8	hamburguesa	0,2476	hamburguesa	0,2636
9	carne	0,2202	té	0,2493
10	pollo	0,2049	café	0,2311
11	**sushi**	0,1954	pollo	0,2064
12	café	0,1722	poroto	0,2040
13	**lechuga**	0,1644	papa frita	0,2020
14	**completo**	0,1554	**lasaña**	0,2012
15	**Pepsi**	0,1477	**puré**	0,1927
16	**Sprite**	0,1446	**lenteja**	0,1868
17	poroto	0,1427	**cazuela**	0,1744
18	**Fanta**	0,1421	**tomate**	0,1679
19	pan	0,1319	pan	0,1578
20	té	0,1258	**papa**	0,1578

En relación con las cinco primeras formas léxicas –sin contar la más disponible– en el listado de EB aparecen, de mayor a menor IDL, *arroz*, *fideo*, *papa frita* y *pizza*; mientras que en el de LH, se leen *agua*, *arroz*, *jugo* y *carne*. Este contraste presenta que, por un lado, en este corte solo coinciden

dos lexías (*Coca-Cola* y *arroz*). Por otro lado, las piezas léxicas de pedagogía –salvo la gaseosa a base de cola– refieren a alimentos sólidos que suelen comerse frecuentemente. En cuanto a las lexías de LH, se combinan comestibles sólidos y líquidos, los que también suelen ser recurrentes en los platos diarios chilenos.

Sobre los veinte vocablos más disponibles, ambas carreras confluyen en catorce de ellos, lo que significa que cada sociolecto exhibe seis palabras distintas, a saber: *sushi* (rango 11), *lechuga* (13), *completo* (14), *Pepsi* (15), *Sprite* (16) y *Fanta* (18), en el diccionario de Educación Básica. Y, *lasaña* (14), *puré* (15), *lenteja* (16), *cazuela* (17), *tomate* (18) y *papa* (20), en el de Letras Hispánicas. Estos ejemplos exponen solamente elemento ligados directamente con el centro de interés, no se hallan entre ellos indicios léxicos de asociaciones indirectas. Sin embargo, debe recalcarse la presencia de: i) marcas comerciales de bebestibles; ii) platillos internacionales; y iii) dialectalismos acerca de alimentos nacionales.

5.2.3. Variable año de curso

Las pruebas estadísticas correspondientes a la variable *año de la carrera* –que está conformada por las variantes 1.er y 4.º año– mostraron que solamente tres de los ochos centros de interés resultaron significativos, a saber: i) *El profesor*, ii) *La escuela: muebles y materiales*, de la muestra de Educación Básica; y iii) *Comidas y bebidas*, del corpus digital de Letras Hispánicas. En virtud de lo cual, en los siguientes párrafos, se exponen los análisis cualitativos de estos actualizadores.

El profesor

Los vocablos *enseñar* –cuyo IDL es igual a 0,2825, en el sociolecto de recién ingreso a la universidad– y *docente* –con un IDL igual a 0,2630 en el más avanzado– son los que encabezan los listados referentes al eje temático *el profesor* de los encuestados de Educación Básica, en virtud de lo cual son los más disponibles y, en consecuencia, corresponden a los elementos nucleares del actualizador. Al otro extremo, en el último puesto de este corte de veinte unidades, se localizan *respeto* (IDL = 0,0832; en 1.er) y *clase* (IDL, 0,0770; en 4.º año). Es decir, los índices empiezan ascendentemente con valores por encima de 0,05 y son más altos –tanto en el mayor como en el menor rango– en el grupo de primer nivel. En la tabla 116, se detallan estos resultados.

Tabla 116. Los 20 vocablos más disponibles del CI *El profesor*, Año de curso

	1.er Año		4.o Año	
	Vocablo	IDL	Vocablo	IDL
1	**enseñar**	0,2825	**docente**	0,2630
2	vocación	0,2226	guía	0,1799
3	**educador**	0,1805	**mediador**	0,1766
4	guía	0,1578	vocación	0,1697
5	conocimiento	0,1534	**responsabilidad**	0,1532
6	paciencia	0,1427	colegio	0,1397
7	colegio	0,1271	conocimiento	0,1323
8	enseñanza	0,1263	**profesional**	0,1127
9	**alumno**	0,1215	educación	0,1080
10	**buena onda**	0,1176	**persona**	0,1064
11	**aprender**	0,1116	**maestro**	0,1057
12	**apoyo**	0,1111	aprendizaje	0,1040
13	**autoridad**	0,1104	paciencia	0,0948
14	educación	0,1024	**pedagogía**	0,0932
15	**sabiduría**	0,0900	**ayuda**	0,0919
16	clase	0,0885	**estudiante**	0,0915
17	aprendizaje	0,0884	enseñanza	0,0897
18	**amor**	0,0884	**docencia**	0,0818
19	**inteligente**	0,0843	**desafío**	0,0782
20	**respeto**	0,0832	clase	0,0770

Respecto a los primeros cinco vocablos, puede observarse que ambos grupos coinciden únicamente en dos de ellos: vocación y guía, en rangos distintos; mientras que los lexemas distintivos de primer año son *enseñar* (rango 1), *educador* (3) y *conocimiento* (5); y los de cuarto año son *docente* (1), *mediador* (3) y *responsabilidad* (5). Así pues, se refleja la alta variabilidad y diferencia entre los dos sociolectos. Al ampliar el espectro y considerar las veinte piezas léxicas, se detalla que los dos grupos coinciden en nueve elementos, mientras que cada uno presenta una suma de once piezas léxicas únicas. Las concernientes al sociolecto de recién ingreso son: *enseñar* (1), *educador* (3), *alumno* (9), *buena onda* (10), *aprender* (11), *apoyo* (12), *autoridad* (13), *sabiduría* (15), *amor* (18), *inteligente* (19) y *respeto* (20). Estas, según la categoría gramatical, son dos verbos, cuyos rasgos léxicos semánticos apuntan a procesos cognitivos; tres adjetivos, que resaltan cualidades básicas y generales; y seis sustantivos, que refieren a un antónimo del nombre del área temática y a percepciones generales de la Figura del educador. Por el contrario, las once palabras exclusivas de la lista de los alumnos aventajados son *docente* (1), *mediador* (3), *responsabilidad* (5), *profesional* (8), *persona* (10), *maestro* (11), *pedagogía* (14), *ayuda* (15), *estudiante* (16), *docencia* (18) y *desafío* (19). Estas pueden clasificarse en cuatro adjetivos y siete sustantivos, las cuales pueden considerarse *grosso modo* con

significados generales; pero ligadas con términos y palabras clave de los requisitos docentes. Así pues, podría indicarse que los lexemas de los participantes de 4.o año parecieran reflejar un poco la metacognición sobre la práctica docente.

La escuela: muebles y materiales

Las palabras con los índices de disponibilidad léxica más altos, en consideración a la variable *año de la carrera*, del centro de interés *La escuela: muebles y materiales*, del alumnado de Educación Básica, son *mesa* y *silla*, con valores iguales a 0,7011 y 0,5118, en 1.er y 4.to año, respectivamente. Así pues, ambos lexemas representan el núcleo léxico del actualizador. Si se centra la atención en la distribución de estas unidades léxicas en las dos listas, se detalla que aparecen una tras otras en orden descendente en el grupo de 1er. año (*mesa* y *silla*), seguidas por *pizarra*. Empero esta acomodación es discontinua en el colectivo de 4.o año, en el que la disposición es *silla*, *pizarra* y *mesa*. Es decir, estos tres lexemas ocupan los tres rangos más altos, a tenor del IDL, pero su ordenamiento es singular en cada grupo, como se detalla en la tabla 117.

Tabla 117. Las 20 palabras más disponibles del CI *La escuela*, según año de curso

Educación Básica				
	1.er Año		4.º Año	
	Vocablo	IDL	Vocablo	IDL
1	mesa	0,7011	silla	0,5118
2	silla	0,6506	pizarra	0,4940
3	pizarra	0,5346	mesa	0,4670
4	lápiz	0,4524	plumón	0,3056
5	plumón	0,3711	sala (de clase)	0,2897
6	cuaderno	0,2925	patio	0,2086
7	**estuche**	0,2622	ventana	0,1875
8	**escritorio**	0,2304	libro	0,1633
9	libro	0,2065	lápiz	0,1576
10	cartulina	0,2060	**data**	0,1341
11	**goma (de borrar)**	0,1924	proyector	0,1288
12	ventana	0,1748	cuaderno	0,1180
13	sala (de clase)	0,1514	cartulina	0,1047
14	**estante**	0,1428	**baño**	0,0995
15	**compu(tador/a)**	0,1426	**biblioteca**	0,0943
16	borrador	0,1301	**lápiz de color**	0,0920
17	patio	0,1245	**cancha (de fútbol)**	0,0881
18	**tijera**	0,1235	**escalera**	0,0868
19	proyector	0,1216	borrador	0,0856
20	**casillero**	0,1188	**antiguo**	0,0763

En conjunto, tanto los discentes de 1er. como los de 4.o año, coinciden en trece de las veinte palabras en las que se ha llevado a cabo el corte. No obstante, difieren en las siguientes siete lexías *estuche* (rango 7), *escritorio* (8), *goma (de borrar)* (11), *estante* (14), *compu(tador/a)* (15), *tijera* (18) y *casillero* (20), por parte del grupo del primer nivel. El grupo del nivel avanzado, por su parte, exhibe los siguientes vocablos distintivos *data* (10), *baño* (14), *biblioteca* (15), *lápiz de color* (16), *cancha (de fútbol)* (17), *escalera* (18) y *antiguo* (20).

Comidas y bebidas

Como se mencionó, respecto a la muestra de Letras Hispánica recogida a través formato digital, solamente resultó significativo el centro de interés *Comidas y bebidas*, correspondiente al factor *año de la carrera*. En la tabla 118, a continuación, se detallan los resultados.

Tabla 118. Las 20 palabras más disponibles del CI *Comidas y bebidas*, según año de curso

Muestra digital: Letras Hispánicas				
	1.er Año		4.o Año	
	Vocablo	IDL	Vocablo	IDL
1	Coca-Cola	0,5329	arroz	0,4295
2	arroz	0,4559	jugo	0,4106
3	fideo	0,3633	Coca-Cola	0,3209
4	jugo	0,3478	carne	0,2738
5	carne	0,3013	**agua**	0,2702
6	té	0,2943	fideo	0,2456
7	pan	0,2593	té	0,2185
8	café	0,2574	hamburguesa	0,2118
9	pizza	0,2406	poroto	0,2088
10	**lechuga**	0,2404	pan	0,2087
11	poroto	0,2394	pizza	0,2066
12	**pollo**	0,2345	**fruta**	0,2036
13	hamburguesa	0,2340	**cazuela**	0,2035
14	**tomate**	0,2248	café	0,2013
15	papa	0,2237	papa	0,1989
16	agua	0,2200	agua	0,1953
17	**Sprite**	0,2089	**verdura**	0,1874
18	**Pepsi**	0,2074	**puré**	0,1837
19	**Fanta**	0,2040	**tallarín**	0,1809
20	**papa frita**	0,2020	**lenteja**	0,1750

En este conjunto, las palabras más disponibles son *Coca-Cola*, cuyo IDL = 0,5329, en el sociolecto de 1.er año; y, *arroz*, con IDL = 0,4295, en el 4.o año. En este sentido, ambos vocablos ocupan los primeros lugares de las listas y, por ende, son los nucleares. Además, debe señalarse que los índices

empiezan, desde el vigésimo hacia el primer puesto, con valores por encima de 0,17; siendo mayor en el grupo de recién ingreso a la universidad, cuyo lexema 20° –*papa frita*– presenta un IDL = 0,2020; mientras que el lexema del grupo avanzado, ubicado en el mismo lugar –*lenteja* (20°)– tiene un valor de IDL = 0,1750. Acerca de las lexías de los cinco primeros lugares, coinciden tres de ellas, salvo *fideo*, en primer año, y *agua*, en cuarto.

Respecto a las veinte palabras, los dos sociolectos coinciden en trece de ellas, pero cada grupo presenta siete únicas. Las siete de los estudiantes de primer año son *lechuga*, que aparece en el rango 10, seguida de *pollo* (12), *tomate* (14), *Sprite* (17), *Pepsi* (18), *Fanta* (19) y *papa frita* (20). En cambio, los siete lexemas exclusivos de 4.º año son *agua*, en el puesto 5; *fruta* (12), *cazuela* (13), *verdura* (17), *puré* (18), *tallarín* (19), *lenteja* (20).

5.2.3. variable formato de las pruebas

En este subapartado se muestran los análisis comparativos de las palabras más disponibles del centro de interés *habilidades y cualidades docentes*, correspondientes a la variable *formato de las pruebas* que, como ya se ha indicado, se ha tomado en cuenta los casos de Letras Hispánicas. La selección de los actualizadores se basó en, primer lugar, a los resultados estadísticos del t-test, según el cual el único eje temático estadísticamente significativo para este grupo fue justamente el CI06.

Habilidades y cualidades docentes

Los vocablos más disponibles en este centro de interés son *paciencia*, en la muestra de Letras Hispánica recolectada de forma digital; este presenta un IDL = 0,4038; y, *empatía*, cuyo IDL es 0,2884, en la muestra de formato tradicional o en papel. Así pues, para el estudiantado de LH la *paciencia* y la *empatía* corresponden los núcleos prototípicos de esta área nocional. Si se observan los valores de los IDL de manera ascendente, podrá detallarse que empiezan por sobre 0,05; específicamente, el lexema *perseverancia* expone un IDL = 0,0538, en el 20.mo lugar de la lista en formato digital; mientras que el lexema *interés* muestra un IDL = 0,0661, en la muestra tradicional, puesto veinte. Además, puede apreciarse, en la tabla 112, que las cinco primeras lexías de ambas listas coinciden, a saber: *paciencia*, *empatía*, *comprensión*, *vocación* y *enseñar*, siendo esta la única que comparte el mismo rango entre los dos grupos.

Tabla 119. Las 20 palabras más disponibles del CI *Habilidades y cualidades docentes*

	Digital		Papel	
	Vocablo	IDL	Vocablo	IDL
1	paciencia	0,4038	empatía	0,2884
2	empatía	0,3910	comprensión	0,2566
3	comprensión	0,2140	vocación	0,2566

4	vocación	0,2018	paciencia	0,2539
5	enseñar	0,1224	enseñar	0,2193
6	**responsabilidad**	0,1190	amabilidad	0,1553
7	**escuchar**	0,1130	conocimiento	0,1552
8	amabilidad	0,1089	respeto	0,1368
9	respeto	0,1064	explicar	0,1226
10	conocimiento	0,0945	sabiduría	0,1081
11	**dedicación**	0,0789	**pedagogía**	0,0920
12	**tolerancia**	0,0689	inteligencia	0,0895
13	inteligencia	0,0678	didáctico	0,0791
14	didáctico	0,0664	**voz**	0,0695
15	pasión	0,0654	pasión	0,0682
16	comunicación	0,0643	**disposición**	0,0672
17	**amor**	0,0640	**estudio**	0,0669
18	sabiduría	0,0628	**hablar**	0,0669
19	explicar	0,0591	comunicación	0,0664
20	**perseverancia**	0,0538	**interés**	0,0661

Conforme a los veinte vocablos más disponibles, puede apreciarse que existe una alta coincidencia entre los dos formatos de toma de datos; pues, tienen una coincidencia de catorce piezas léxicas. Por el contrario, cada conjunto muestra una diferencia de seis palabras. En el caso de la muestra digital, las seis palabras únicas son *responsabilidad*, rango 6; *escuchar* (7), *dedicación* (11), *tolerancia* (12), *amor* (17) y *perseverancia* (20); mientras que las seis exclusivas de la muestra en papel son *pedagogía*, en el rango 11; *voz*, en el 14; *disposición* (16), *estudio* (17), *hablar* (18) e *interés* (20).

5.2.4. Variable Frecuencia de lectura

Sobre la variable *Frecuencia de lectura optativa*, la cual está compuesta por cuatro variantes, la prueba Anova arrojó como significativos los centros de interés *La lectura*, *La educación*, *Juegos y distracciones* y *Habilidades y cualidades docentes*, de la muestra digital del grupo de Letras Hispánicas.

La lectura

De las veinte palabras más disponibles del CI01, *libro* es la que ocupa el primer lugar en las cuatro variantes de *frecuencia de lectura optativa*, por lo que puede entenderse como el elemento léxico nuclear del área nocional. El cómputo más elevado de este lema se halla en el último nivel de análisis (Más de 10 horas), IDL = 1,0000; opuestamente, el menor valor se aprecia en el primer nivel (Ninguna hora), con un IDL = 0,7978. A *libro* le siguen los vocablos *letras* –que ocupa el segundo lugar en las variantes: Ninguna, De 1 a 5, y Más de 10 horas, con el valor más alto en la primera

(IDL = 0,3852)– y *palabra* –esta ocupa la segunda posición en De 6 a 10 horas (IDL = 0,3553) y el tercer puesto en las dos primeras variantes–. En cuanto a las listas de palabras desde el vigésimo al primer rango, se detalla que los lemas que se localizan en el último rango son: *oración* (IDL = 0,0717), *conocimiento* (IDL = 0,1035), *poesía* (IDL = 0,0824) y *palabra* (IDL = 0,1200), según se leen en la tabla 120.

En esta misma línea argumental, entre las cinco palabras más disponibles, se aprecia que las dos primeras variantes comparten los mismos cinco vocablos, los que, además, coinciden en sus posiciones respecto al rango. Estos lexemas –organizados de manera descendente– son: *libro*, *letra*, *palabra*, *autor* y *leer*. Contrariamente, la tercera variante presenta una disimilitud con respecto a los otros niveles de la variable, ya que posee una lexía distintiva, a saber: *comprensión (lectora)*, la cual no se comparte con ninguno, al menos en este primer corte. Por su parte, en la última variante, son dos los lemas que se diferencian del resto, estos son: *novela* y *entretención*. Es decir, el nivel que describe las prácticas lectoras como la de mayor frecuencia o hábito, es la que ha resultado con un cómputo mayor de piezas léxicas, aunque este no sea muy alto.

En consideración al grado de coincidencia léxica, existe una alta sincronía de vocablos en los cuatro niveles, la cual varía desde 16 lemas (ninguna) hasta 13 lexías compartidas (más de 10 horas). Sin embargo, deben resaltarse la disensión observable, la cual va paulatinamente en ascenso a medida que se pasa de la primera a la última variante. Así pues, se evidencia que las variantes 1 y 2 tienen el menor número de vocablos exclusivos; específicamente, la primera tiene 4 piezas léxicas únicas: *hábito*, rango 12; *ensayo (académico)*, rango 14; *comprender* (18) y *oración* (20); mientras que la segunda posee 5: *página* (7), *lápiz* (10), *imaginación* (14), *personaje* (16) y *párrafo* (18). Por su parte, la tercera y cuarta variante tienen más unidades léxicas divergentes. Concretamente: en la tercera, los vocablos únicos son: *ficción* (8), *aprender* (10), *hoja* (13), *revista* (16), *práctica* (17), *poema* (18); en la cuarta, se aprecian siete: *entretención*, puesto 5; *saga* (10), *librería* (11), *paper* (12), *educación* (15), *texto crítico* (18), *IVA* (19).

La educación

La palabra más disponible del centro de interés *la educación*, en cuanto a los resultados de la variable *frecuencia de lectura*, es *profesor*, que ocupa el primer puesto en las listas de las tres primeras variantes (ninguna, de 1 a 5 y de 6 a 10 horas), mostrando el valor más alto en la segunda variante (IDL = 0,4927). Sin embargo, en el último nivel de análisis, el lexema *derecho* es el que va a la cabeza de la lista, siendo, por lo tanto, el más disponible para este grupo, con un IDL = 0,3716. No obstante,

debe notarse que el lema *profesor*, si bien sí se encuentra en el listado de este nivel; se localiza en el segundo lugar, con un IDL = 0,3716. Así pues, los vocablos *profesor* y *derecho* resultan ser los elementos léxicos prototípico del actualizador *la educación*. En esta área nocional, al menos, para los participantes que se decantaron por las variantes: ninguna, de 1 a 5 y 6 de 10 horas. Como se detallan en la tabla 121.

Juegos y distracciones

Los índices de disponibilidad léxica indican que –para el área nocional *juegos y distracciones*, en relación con la variable *frecuencia de lectura*– el vocablo *diversión* es el más disponible y, por ende, el más nuclear en la primera y tercera variantes, con valores iguales a IDL = 0,2687 y IDL = 0,3485, respectivamente. En cambio, en la segunda y cuarta variante, los lexemas más disponibles son *amigo* (IDL = 0,1867) y *entretención* (IDL = 0,6679), respectivamente. Al detallar las primeras cinco piezas léxicas con los valores de disponibilidad más altos, puede apreciarse que existe una gran variabilidad en el vocabulario de este campo nocional, puesto que el grado de coincidencia entre los niveles de análisis es muy poco. En concreto, solo los lemas *diversión* y *amigo* se comparten en los tres primeros niveles. Además, cada uno de ellos exhibe tres unidades léxicas diferentes. En relación con el cuarto nivel, las cinco piezas léxicas más disponibles son exclusivas en contraste con las otras tres variantes, según se lee en la Tabla 122.

Sobre los índices, se aprecia que los valores de los veinte vocablos más disponibles difieren en cada variante. Así pues, el nivel que supone una mayor práctica lectora por parte de los encuestados presenta el mayor número de piezas léxicas con IDL≥0,1 (todas las palabras del corte). En este caso, la lexía de rango 20 (*música*) tiene IDL = 0,1263 y la de rango 1 (*entretención*), IDL = 0,6679. Por su parte, el tercer nivel cuenta con diecinueve lexías con IDL≥0,1; mientras que el primero posee diecisiete y el segundo nivel, trece.

Tabla 120. Las 20 palabras más disponibles del CI *La lectura*, en la muestra digital

	Ninguna		1 a 5		6 a 10		Más de 10	
	Vocablo	IDL	Vocablo	IDL	Vocablo	IDL	Vocablo	IDL
1	libro	0,7978	libro	0,9150	libro	0,8462	libro	1,0000
2	letra	0,3852	letra	0,3586	palabra	0,3553	letra	0,3520
3	palabra	0,2083	palabra	0,2526	letra	0,3181	novela	0,2749
4	autor	0,1913	autor	0,2115	comprensión (lectora)	0,2636	leer	0,2605
5	leer	0,1532	leer	0,1984	autor	0,2245	**entretención**	0,2564
6	papel	0,1320	novela	0,1770	leer	0,2002	comprensión (lectora)	0,2563
7	cuento	0,1197	**página**	0,1708	biblioteca	0,1786	escribir	0,2478
8	escritura	0,1114	literatura	0,1686	**ficción**	0,1737	cuento	0,2333
9	escritor	0,1097	biblioteca	0,1607	escritor	0,1734	literatura	0,2182
10	texto	0,1068	**lápiz**	0,1351	**aprender**	0,1556	**saga**	0,1920
11	escribir	0,1054	lentes	0,1301	literatura	0,1550	biblioteca	0,1760
12	**hábito**	0,1009	escritura	0,1259	cuento	0,1430	**librería**	0,1760
13	conocimiento	0,1001	texto	0,1256	**hoja**	0,1382	**paper**	0,1760
14	**ensayo (académico)**	0,0987	**imaginación**	0,1245	lentes	0,1197	conocimiento	0,1549
15	poesía	0,0983	cuento	0,1234	cultura	0,1131	**educación**	0,1420
16	comprensión (lectora)	0,0958	**personaje**	0,1173	**revista**	0,1120	poesía	0,1375
17	literatura	0,0898	comprensión (lectora)	0,1166	**práctica**	0,1120	texto	0,1363
18	**comprender**	0,0815	**párrafo**	0,1143	**poema**	0,1042	**texto crítico**	0,1363
19	cultura	0,0728	escribir	0,1123	papel	0,0997	**IVA**	0,1363
20	**oración**	0,0717	conocimiento	0,1035	poesía	0,0824	palabra	0,1200

Tabla 121. Las 20 palabras más disponibles del CI *La educación*, en la muestra digital

	Ninguna		1 a 5		6 a 10		Más de 10	
	Vocablo	IDL	Vocablo	IDL	Vocablo	IDL	Vocablo	IDL
1	profesor	0,4109	Profesor	0,4946	profesor	0,4027	derecho	0,3716
2	conocimiento	0,3490	colegio	0,2167	alumno	0,3055	profesor	0,3516
3	**formación**	0,2688	libro	0,2067	conocimiento	0,3043	enseñanza	0,2684
4	aprendizaje	0,2542	aprendizaje	0,2029	universidad	0,1961	alumno	0,2192
5	colegio	0,2358	derecho	0,1966	enseñanza	0,1670	aprendizaje	0,2058
6	importante	0,2247	aprender	0,1889	colegio	0,1622	**esencial**	0,2000
7	cuaderno	0,2193	**necesario**	0,1862	aprendizaje	0,1618	**desigual**	0,2000
8	universidad	0,1736	conocimiento	0,1848	**información**	0,1569	colegio	0,1988
9	alumno	0,1616	alumno	0,1778	**matemáticas**	0,1487	calidad	0,1975
10	libro	0,1354	estudiante	0,1613	**lenguaje**	0,1470	vocación	0,1946
11	lápiz	0,1293	importante	0,1306	**sala (de clases)**	0,1447	estudiante	0,1880
12	enseñanza	0,1288	universidad	0,1243	**lucro**	0,1334	conocimiento	0,1880
13	**estudio**	0,1247	futuro	0,1164	deficiente	0,1294	**cultura**	0,1716
14	**experiencia**	0,1247	lápiz	0,1125	aprender	0,1278	**malo**	0,1716
15	deficiente	0,1202	cuaderno	0,1028	libro	0,1275	**escolar**	0,1716
16	**proceso**	0,1163	**necesidad**	0,0923	calidad	0,1218	universidad	0,1706
17	**estudiar**	0,1163	**escuela**	0,0740	**liceo**	0,1213	**transmisión (de conocimiento)**	0,1586
18	futuro	0,1084	**sociedad**	0,0713	**estrategia**	0,1175	**acceso**	0,1472
19	aprender	0,1014	**familia**	0,0705	**enseñar**	0,1145	**educación universitaria**	0,1472
20	vocación	0,0976	enseñanza	0,0690	**modales**	0,1093	libro	0,1263

Continuando con el análisis del CI04, las convergencias y divergencias léxicas de las cuatro variantes de *frecuencia de lectura* denotan que a medida que se pasa de un nivel inferior (Ninguna hora) a uno superior (Más de 10 horas), el número de piezas léxicas únicas aumenta. En correspondencia con lo anterior, se aprecia que el primer nivel tiene una coincidencia de dieciséis palabras con los otros niveles, empero se distingue de ellos con cuatro, a saber: *necesidad*, en el rango 8, *ludo* (16), *salir* (17) y *pelota* (19). Por su parte, el segundo nivel concierta en quince unidades léxicas con el resto de los niveles, pero se destaca de ellos en cinco: *YouTube* (9), *libro* (12), *correr* (17), *pintar* (18) y *necesario* (20). Por su lado, el tercer nivel ofrece una confluencia de catorce palabras y, por lo tanto, una discrepancia de seis: *aprender* (4), *Catan* (6), *deporte* (8), *infancia* (15), *lectura* (16) y *Dixit* (20). Por último, el cuarto nivel de análisis expone una concurrencia de nueve lexías, mientras que se separa de las demás variantes por contener once unidades léxicas diferentes: *recreación* (3), *relajación* (4), *competencia* (5), *entretenimiento* (7), *aire libre* (11), *disfrute* (13), *dominó* (14), *salud (mental)* (15), *estrés* (16), *Play(Station)* (17) y *felicidad* (18).

Habilidades y cualidades docentes

El último centro de interés que resultó estadísticamente significativo a través de los cálculos de Anova respecto a la variable *frecuencia de lectura* fue *habilidades y cualidades docente* de la muestra digital de Letras Hispánicas. Así pues, al seleccionar los veinte vocablos más disponibles de este actualizador, se observa que los que ocupan el rango 1 de los listados son: *empatía* y *paciencia*. El primero encabeza los listados de las variantes 1 y 4, mostrando el IDL más alto en el nivel Ninguna, con 0,4823; mientras que en Más de 10 hora alcanzó un IDL = 0,3589. En cuanto al lema *paciencia*, este se ubica en el primer rango de la 2.ª y 3.ª variante, con IDL = 0,4130 y 0,5234, respectivamente. En este contexto, puede señalarse que ambas palabras representan el núcleo léxico de los vocabularios analizados.

Si bien se han contrapuesto las veinte palabras con los mayores índices de disponibilidad por variantes, se ha empezado con el contraste de las cinco primeras –como se ha efectuado con los demás ejes temáticos–. En consonancia con lo anterior, esta primera exploración ha permitido determinar que los lexemas *empatía*, *paciencia* y *comprensión* son los que se repiten en los cuatro niveles de análisis. No obstante, las variantes 1, 2 y 3 exhiben una concurrencia de cuatro lemas; es decir, además de los tres previamente mencionados, se encuentra también *vocación*. Así pues, en este corte, estas tres variantes ostentan una palabra diferente cada una. Concretamente, el nivel "ninguna" presenta el vocablo *enseñar*; mientras que "de 1 a 5 horas" y "de 6 a 10 horas" introducen los lexemas

responsabilidad y *escuchar*, respectivamente. Por su parte, la variante 4 ostenta dos lexías exclusivas, a saber: *conocimiento* y *transmitir conocimiento*. Estos datos se ilustran en la tabla 123.

Tabla 122. Comparación de las cinco palabras más disponibles del CI: *la lectura*

	Ninguna	De 1 a 5 horas	De 6 a 10 horas	Más de 10 horas
1	empatía	paciencia	paciencia	empatía
2	vocación	empatía	empatía	Paciencia
3	paciencia	comprensión	**escuchar**	**conocimiento**
4	comprensión	**responsabilidad**	comprensión	**transmitir conocimiento**
5	**enseñar**	vocación	vocación	comprensión

En cuanto a las veinte lexías más disponibles (Tabla 124), los resultados patentizan que la variante que representa la frecuencia de lectura más alta (más de 10 horas) cuenta con el mayor número de lexías distintivas en relación con las otras tres, exactamente doce piezas léxicas. Estas –ordenadas de mayor al menor rango– son: *transmitir conocimiento* (4), *aprendizaje* (6), *moderación* (9), *enseñanza* (10), *educación* (11), *firmeza* (12), *herramienta (pedagógica)* (13), *entendimiento* (14), *autoridad* (15), *conciencia* (17), *crítico* (18) y *evaluación* (20). En segundo lugar, según la suma de lexías distintivas, se encuentra la variante 1 (Ninguna), con nueve lexías, a saber: *perseverancia* (10), *entretenido* (11), *profesión* (12), *bailar* (14), *competencia (cognitiva/ cultural)* (15), *respetuoso* (16), dinamismo (17), puntualidad (19) y cantar (20). En cuanto a las variantes 2 (De 1 a 5) y 3 (De 6 a 10), ambas presentan el mismo cómputo de palabras diferentes, particularmente, ocho. Las piezas léxicas exclusivas de la 2ª variante son: *dedicación* (8), *inteligencia* (10), *sabiduría* (11), *compromiso* (13), *pasión* (15), *atención* (17), *entender* (18) e *interés* (20); mientras que los ocho vocablos distintivos de la 3ª variante son: *amor* (8), *cariño* (9), *comprensivo* (13), *simpatía* (14), *voluntad* (16), *disposición* (17), *motivación* (19) y *hablar* (20).

Tabla 123. Las 20 palabras más disponibles del CI *Juegos y distracciones*

	Ninguna		1 a 5		6 a 10		Más de 10	
	Vocablo	IDL	Vocablo	IDL	Vocablo	IDL	Vocablo	IDL
1	diversión	0,2687	amigo	0,1867	diversión	0,3485	entretención	0,6679
2	amigo	0,2352	leer	0,1556	videojuego	0,3105	online	0,3783
3	tiempo libre	0,2211	diversión	0,1433	amigo	0,2722	**recreación**	0,2679
4	descanso	0,2020	película	0,1341	**aprender**	0,2131	**relajación**	0,2092
5	celular	0,1946	compu(tador/a)	0,1331	ocio	0,1782	**competencia**	0,2050
6	película	0,1839	videojuego	0,1297	**Catan**	0,1755	videojuego	0,2038
7	internet	0,1739	serie	0,1269	pasatiempo	0,1605	**entretenimiento**	0,2000
8	**necesidad**	0,1351	internet	0,1253	**deporte**	0,1594	juego de mesa	0,2000
9	jugar	0,1313	**YouTube**	0,1131	película	0,1578	carta	0,1783
10	videojuego	0,1267	entretención	0,1125	música	0,1546	internet	0,1783
11	pasatiempo	0,1239	música	0,1044	serie	0,1544	**aire libre**	0,1783
12	hobby	0,1185	**libro**	0,1034	leer	0,1512	descanso	0,1783
13	compu(tador/a)	0,1135	celular	0,1031	hobby	0,1426	**disfrute**	0,1589
14	carta	0,1095	carta	0,0962	tiempo libre	0,1292	**dominó**	0,1589
15	ocio	0,1088	Monopoly	0,0951	**infancia**	0,1276	**salud (mental)**	0,1589
16	**ludo**	0,1062	niño	0,0930	**lectura**	0,1252	**estrés**	0,1416
17	**salir**	0,1054	**correr**	0,0896	juego de mesa	0,1163	**Play(Station)**	0,1416
18	leer	0,0795	**pintar**	0,0866	Monopoly	0,1094	**felicidad**	0,1263
19	**pelota**	0,0734	juego de mesa	0,0858	jugar	0,1001	niño	0,1263
20	online	0,0713	**necesario**	0,0830	**Dixit**	0,0960	música	0,1263

Tabla 124. Las 20 palabras más disponibles del CI *Habilidades y cualidades docentes*

	Ninguna		1 a 5		6 a 10		Más de 10	
	Vocablo	IDL	Vocablo	IDL	Vocablo	IDL	Vocablo	IDL
1	empatía	0,4823	paciencia	0,4130	paciencia	0,5234	empatía	0,3589
2	vocación	0,3728	empatía	0,3716	empatía	0,3036	paciencia	0,2633
3	paciencia	0,2463	comprensión	0,1855	escuchar	0,2879	conocimiento	0,2633
4	comprensión	0,2192	responsabilidad	0,1482	comprensión	0,2638	**transmitir conocimiento**	0,2000
5	enseñar	0,1747	vocación	0,1396	vocación	0,1956	**compresión**	0,2000
6	didáctico	0,1226	amabilidad	0,1298	enseñar	0,1638	**aprendizaje**	0,1589
7	responsabilidad	0,1059	respeto	0,1188	tolerancia	0,1557	respeto	0,1589
8	comunicación	0,1022	**dedicación**	0,1003	**amor**	0,1554	enseñar	0,1589
9	aprender	0,0797	escuchar	0,0997	**cariño**	0,1302	**moderación**	0,1589
10	**perseverancia**	0,0720	**inteligencia**	0,0910	explicar	0,1255	**enseñanza**	0,1263
11	**entretenido**	0,0699	**sabiduría**	0,0855	respeto	0,1128	**educación**	0,1263
12	**profesión**	0,0667	conocimiento	0,0837	aprender	0,1091	**firmeza**	0,1263
13	explicar	0,0667	**compromiso**	0,0830	**comprensivo**	0,0929	**herramienta pedagógica**	0,1263
14	**bailar**	0,0667	enseñar	0,0816	**simpatía**	0,0920	**entendimiento**	0,1003
15	**competencia**	0,0667	**pasión**	0,0792	psicología	0,0920	**autoridad**	0,1003
16	**respetuoso**	0,0667	tolerancia	0,0725	**voluntad**	0,0890	psicología	0,1003
17	**dinamismo**	0,0667	**atención**	0,0715	**disposición**	0,0878	**conciencia**	0,1003
18	amabilidad	0,0630	**entender**	0,0710	didáctico	0,0873	**crítico**	0,0797
19	**puntualidad**	0,0559	comunicación	0,0648	**motivación**	0,0871	vocación	0,0797
20	**cantar**	0,0559	**interés**	0,0584	**hablar**	0,0839	**evaluación**	0,0703

5.3. Comparación intermuestral del léxico disponible

En este epígrafe se comparan las diez palabras más disponibles de los centros de interés que, cuantitativamente, resultaron más productivos y compactos, a tenor de los promedios de palabras e índice de cohesión, a saber: 1) *partes del* cuerpo, 2) *Comidas y bebidas* y 3) *La escuela: muebles y materiales*, de las tres muestras de este estudio con las de los trabajos de Valencia y Echeverría (1999), Herranz (2020) y Martínez-Lara (2021), por las siguientes razones:

- Valencia y Echeverría (1999), *Léxico disponible de Chile*, analizaron el LD de 600 alumnos de secundaria, de distintas localidades chilenas, desde el norte hasta el sur del país. Plantearon 18 centros de interés, algunos de ellos semejantes a los del PPHLD.
- Herranz (2020), *Palabra de maestros. Análisis del léxico disponible de los futuros docentes*, desarrolló su investigación con base en las listas de palabras elaboradas por 591 universitarios españoles, de 1.er y 4.o año de curso, de las carreras de Educación Infantil y Educación Primaria, de la Universidad Rey Juan Carlos, Madrid. La autora testeó a los informantes sobre los dieciséis CI del PPHLD más dos nuevos, a saber: (17) *Nuevas tecnologías: TIC* y (18) *Educación*.
- Martínez-Lara (2021), *Incidencia de los años de escolaridad y cantidad de lectura en el léxico disponible de un grupo de estudiantes universitarios del área de pedagogía*, se basó en las respuestas de 57 discentes, de recién ingreso y último año, de carreras del área de pedagogía: Castellano y Filosofía, Ciencias biológicas y Parvularia, de la Universidad de La Serena, Chile. Los ejes temáticos explorados fueron *La escuela: muebles y materiales*, *la educación* y *la lectura*.

En virtud de que estas investigaciones no comparten a cabalidad las mismas categorías semánticas, el análisis comparativo se efectuó con las áreas nocionales coincidentes.

La escuela: muebles y materiales

Por una parte, se observa, en la Tabla 125, que las palabras que encabezan las listas y, por ende, las más disponibles del CI05 son *mesa* y *silla*; estas representan, entonces, los núcleos léxicos de esta categoría, como se lee en la tabla 118. Por otra parte, las tres muestras de esta tesis tienen una alta coincidencia léxica, pero se distinguen del trabajo de Herranz (2020), por cuatro unidades léxicas, a saber: *bolí(grafo)*, *tiza*, *rotulador* y *pupitre*. De estas debe resaltarse *bolí(grafo)*, que, si bien está

registrada en el DLE (2014) sin marca dialectal, parece tener más uso en España[20]. Contrariamente, en Chile, tiene una baja frecuencia, al menos, así parece evidenciarlo su ausencia en el diccionario de este trabajo y en el de Martínez-Lara (2021).

Tabla 125. Comparación intermuestral del léxico disponible del CI *La escuela*

EB	LH2	LH (digital)	Herranz (2020)	Martínez-Lara (2021)
mesa	mesa	mesa	silla	mesa
silla	silla	silla	mesa	silla
pizarra	lápiz	lápiz	pizarra	pizarra
lápiz	pizarra	pizarra	**bolí(grafo)**	plumón
plumón	cuaderno	cuaderno	lápiz	**lápiz (de color)**
cuaderno	plumón	libro	**tiza**	cuaderno
estuche	estuche	plumón	**rotulador**	libro
libro	goma (de borrar)	estuche	libro	**computador**
escritorio	libro	escritorio	**pupitre**	**estante**
sala (de clases)	**borrador**	goma (de borrar)	cuaderno	estuche

Partes del cuerpo

Respecto al eje temático *Partes del cuerpo*, las diez primeras palabras de los trabajos comparados muestran una total similitud; pues, no se hallan vocablos diferentes entre ellos. De hecho, solo se aprecian disimilitudes en las posiciones que ocupan las palabras en los listados. Entre estas destacan *ojo*, *mano* y *cabeza*, que son las que se localizan en los primeros puestos de los diccionarios confrontados, y *oreja* que cierra cuatro de los cinco vocabularios, como se lee en la Tabla 126.

Tabla 126. Comparación intermuestral del léxico más disponible del CI *Partes del cuerpo*

EB (en papel)	LH (en papel)	LH (digital)	Valencia y Echeverría (1999)	Herranz (2020)
mano	ojo	ojo	cabeza	ojo
ojo	mano	mano	brazo	cabeza
cabeza	nariz	cabeza	pierna	brazo
pierna	pie	pie	mano	nariz
brazo	cabeza	nariz	ojo	mano
pie	brazo	brazo	dedo	pierna
dedo	dedo	pierna	pie	dedo
nariz	pierna	dedo	nariz	boca
uña	uña	oreja	boca	pie
oreja	oreja	boca	oreja	oreja

[20] Al respecto, en las investigaciones de Hernández Muñoz (2004) y Santos Díaz (2020) también se encuentra entre las diez primeras palabras.

Comidas y bebidas

Por último, el área nocional referida a los alimentos y las bebidas presenta divergencias léxicas de carácter, mayormente, dialectal, puesto que los dos trabajos españoles registraron los lexemas *zumo* y *patata*, que son las formas más recurrentes en España, mientras que sus contrapartes americanas, *jugo* y *papa*, se encuentran en los datos chilenos. Asimismo, en los diccionarios de Herranz (2020) y Santos-Díaz (2020) se codifica *cerveza*, entre los diez primeros vocablos, que es el lema de uso panhispánico para la bebida alcohólica a base de cebada, como se detalla en la Tabla 127.

Tabla 127. Comparación intermuestral del léxico más disponible sobre el CI comidas y bebidas

EB	LH1	LH2	Herranz (2020)	Santos-Díaz (2020)
Coca-Cola	Coca-Cola	arroz	agua	leche
arroz	agua	agua	Coca-Cola	agua
fideo	arroz	jugo	pescado	cerveza
papa frita	jugo	Coca-Cola	carne	carne
pizza	carne	fideo	**zumo**	**vino**
jugo	fideo	pan	leche	café
agua	pizza	carne	**Fanta**	pan
hamburguesa	hamburguesa	té	pan	pescado
carne	té	**papa**	**tomate**	**patata**
pollo	café	**lechuga**	cerveza	**manzana**

En síntesis, se observa que los vocabularios referidos a estos tres centros de interés se asemejan bastante, ya que las piezas léxicas exhiben más coincidencias que disimilitudes, sobre todo, en el CI07, en el que las diez palabras más disponibles encajan totalmente. Por su parte, el CI08 refleja mejor la realidad de las variedades del español, concretamente, entre España y Chile.

5.4. Análisis de relaciones asociativas a través de grafos

Algunas de las líneas de investigación de las ciencias cognitivas han tenido interés por los estudios de disponibilidad léxica, ya que gracias a estos pueden explorarse los mecanismos de activación de las palabras. Además, permiten conocer a grandes rasgos los procesos involucrados en la estructuración del lexicón (Henríquez *et al.*, 2016: 230). De esta corriente, los lexicógrafos han tomado en cuenta, habitualmente, las teorías de: i) prototipos (Rosch, 1973, 1975; Rosch y Mervis, 1975; Cuenca y Hilferty, 1999; entre otros), ii) semántica cognitiva (Gibbs, 1996; Muñoz, 2006; Valenzuela, Ibarretxe-Antuñano y Hilferty, 2012) y iii) redes semánticas (Echeverría y Ferreira, 2010; Hernández Muñoz, 2006; Echeverría *et al.*, 2008; Manjón-Cabeza, 2008, 2010; Santos Díaz, 2017c).

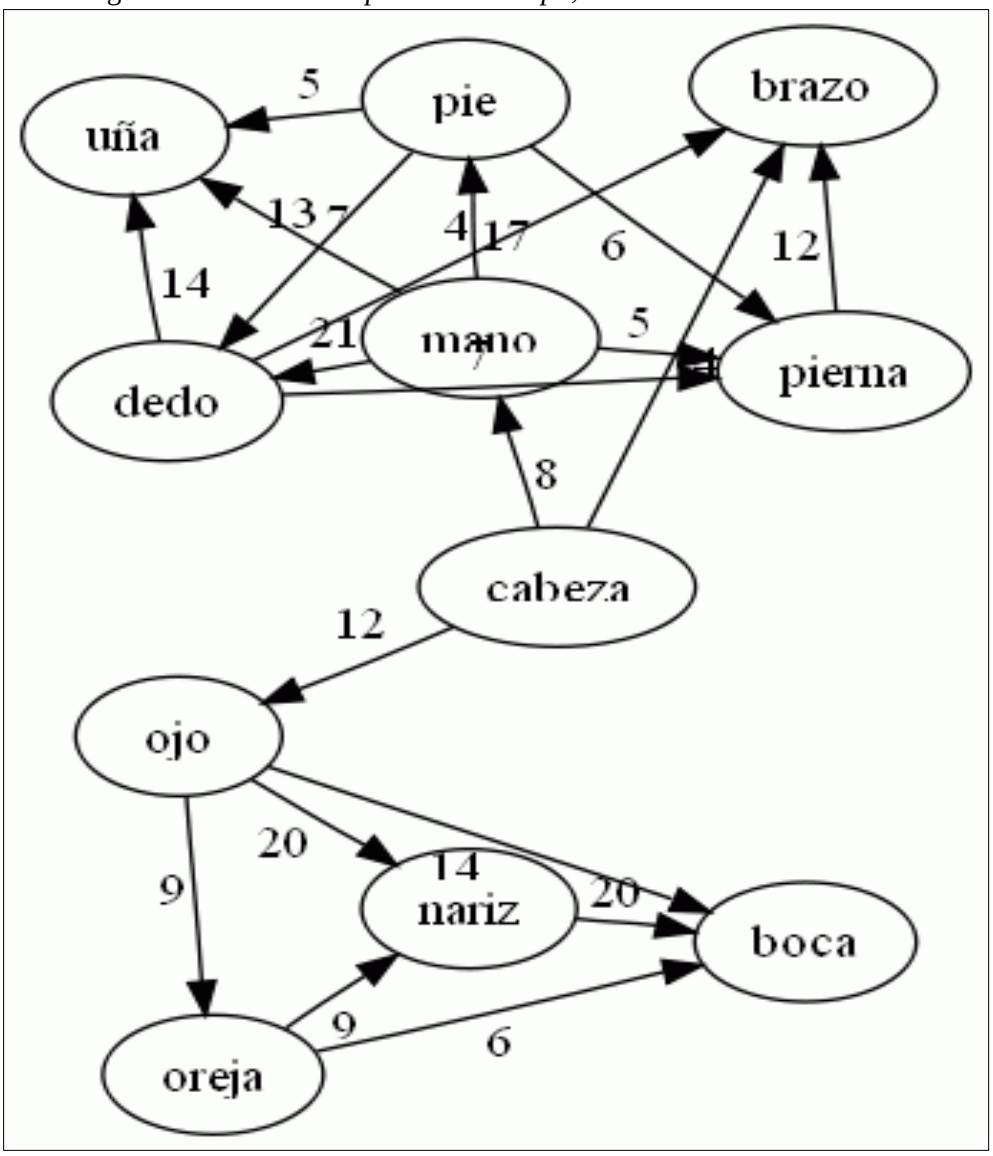
Respecto a esta última, las pesquisas se han llevado a cabo a través del análisis de grafos, los cuales son elaborados mediante programas computacionales, como Dispografo (Echeverría *et al.*, 2008).

En los trabajos referidos a los análisis de grafos, como los de Ferreira y Echeverría (2010), Henríquez *et al.* (2016); Hernández Muñoz y Tomé (2017), Mahecha y Mateus-Ferro (2017); Mateus-Ferro *et al.* (2018), se ha buscado describir y explicar los mecanismos psicolingüísticos que podrían motivar las conexiones de los vocablos más disponibles. Pues, los grafos muestran el tejido de la red en el que se hallan las palabras, exhibiendo los enlaces entre ellas. Al respecto, HHernández-Muñoz y Tomé (2017: 111) aseveran que "Las representaciones de grafos muestran redes de nodos conectados a través de aristas (los grafos), que se interpretan como redes semánticas cuya configuración expresa las relaciones subyacentes en los corpus de disponibilidad. Los nodos son los vocablos disponibles y las aristas expresan las relaciones entre nodos". De esta manera, los grafos permiten observar i) los lazos que tiene una lexía X con otras, ii) el vecindario en el que se sitúa, ii) la intensidad asociativa que dicho lexema tiene con los otros, pero también propicia la visualización de los elementos desconectados y, por ende, aislados del resto.

En virtud de lo anterior y consecuente con los objetivos de esta tesis, se analizaron los grafos del centro de interés *partes del cuerpo*, ya que este resultó ser el más productivo y compacto de los ocho propuestos, en las tres muestras. La indagación se basó en la identificación de los nodos conglomerados, a razón de sus asociaciones semánticas.

En la figura 28, se aprecia la red asociativa de las listas de palabras de los estudiantes de Educación Básica, resultante de la limpieza del grafo general por medio de la poda de las aristas ≤ 3 y de los nodos ≤ 2. En este pueden detallarse, claramente, dos grandes grupos semánticos vinculados a través del nodo *cabeza*, que funciona como núcleo de la red. Por un lado, se encuentra el conjunto semántico referido a las extremidades (superiores e inferiores) del cuerpo humano; mientras que, por otro lado, se halla el colectivo léxico que remite a los órganos sensoriales localizados en la cabeza. Respecto a las extremidades, la red expone una asociación directa entre *cabeza*, *brazo* y *mano*, siendo esta el centro del subgrupo. En este se exhiben relaciones semánticas de holónimo con sus respectivos merónimos: *mano-dedo-uña*, en la que *mano* es el holónimo. También, hay conexiones basadas en afinidad semántica: *mano-pie*, *pierna-brazo*, ya que una cumple la función (a grandes rasgos) de la otra; así pues, la mano (donde están los dedos) es la parte terminal del brazo, mientras que el pie es la de la pierna. Por otro lado, resulta interesante que –si bien los órganos sensoriales de este subgrupo (*ojo*, *nariz*, *boca* y *oreja*) se localizan en la cabeza– el nodo de esta se liga directa y únicamente con el nodo *ojo*. En este se aprecia solo asociaciones semánticas de afinidad y proximidad visual.

Figura 28. Grafo del CI *partes del cuerpo*, de la muestra de Ed. Básica



En cuanto a la red asociativa sobre la fisionomía humana trazada a partir de los datos de los alumnos de Letras Hispánicas que realizaron las pruebas en papel, se observan dos grandes grupos inconexos. De un lado, se agrupan cuatro nodos, creando un bucle casi cerrado, separados del resto. Esta agrupación concierne a los órganos internos (*pulmón*, *corazón* y *riño*), cuya asociación se debe justamente a sus afinidades semánticas; también aparece *costilla*, que en este caso guarda un lazo con los otros debido a la proximidad y ubicación visual. Pues, tanto el corazón como los pulmones están juntos dentro de las costillas, en la parte superior; de modo que estas funcionan como un muro defensor; mientras que los riñones, siendo órganos también, quedan un poco más debajo de las costillas y, por lo tanto, más alejado, como se aprecia en la figura 29.

Más complejo es el segundo subgrupo, ya que está integrado por un número mayor de nodos. En este se captan dos subconjuntos de palabras: 1) las relativas a la parte superior del cuerpo, con especial énfasis en la cabeza; y 2) las atinentes a las extremidades, superiores e inferiores. Con relación al primero, el núcleo es *ojo*, y se muestran encadenamientos semánticos por afinidad y proximidad visual: *ojo-nariz-boca-oreja*; *ceja-pelo*; *pelo-cabeza-cuello*; empero también hay una conexión por holonimia: *ojo-ceja*. Sobre el segundo subgrupo, el centro es *mano*, que expone nexos por holonimia y meronimia: *mano-dedo-uña*, al igual que *pierna-rodilla-pie-dedo-uña*; existen también uniones por afinidad semántica y proximidad visual: *mano-pie*, *rodilla-codo*, *pierna-brazo*.

Con base en los datos de la muestra digital, se ha reproducido un grafo del eje temático *Partes del cuerpo*, ilustrado en la imagen 30. En este se aprecian tres conjuntos semánticos facilmente deducibles a la vista, uno de los cuales –al igual que ocurrió con los datos de la muestra 2– se emplaza separado, como una isla léxica. Este lo componen los nodos *cerebro*, *corazón* y *pulmón*, que crean un bucle triángular casi cerrardo. Las conexiones entre esta triada se da en función de afinidad semántica, ya que son órganos internos del cuerpo, y por próximidad: *corazón-pulmón*.

En la red más compleja de este sociolecto, pueden identificarse dos conglomerados: 1) vocablos que aluden a las partes de la cabeza, y 2) los que apuntan a las extremidades, ambos unidos por los nodos *ojo* (núcleo) y *cabeza*. Este entramado revela similitudes con los otros dos grafos previamente descritos, ya que exhibe vinculaciones semánticas, tanto por afinidad como por proximidad: *ojo-nariz-boca-oreja*; *mano-pie*; *pierna-brazo*. Empero se distingue de los otros grafos, porque agrega nuevas conexiones de holónimos y merónimos, concretamente: *labio-diente-lengua*; *boca-lengua*; *pierna-muslo-pantorrilla*; *pierna-rodilla-pantorrilla*; asimismo, una conexión por afinidad semántica: *uña-pelo*.

Con más detalles, la agrupación en la que *cabeza* es el núcleo, se observa una expresión ligada a la localización de esta en la fisionomía y a su constitución, más precisamente los nodos *cabeza-cuello-hombro-pelo*, mantienen un enlace de próximidad visual, que puede explicarse como sigue: la *cabeza* se encuentra pegada al *cuello* y este al *hombro*, y ambos unen la cabeza al cuerpo y guardan su estabilidad. Asimismo, se encuentra coronada por *pelo*.

Figura 29. Grafo del CI07 de la muestra de Letras H. recogida en papel

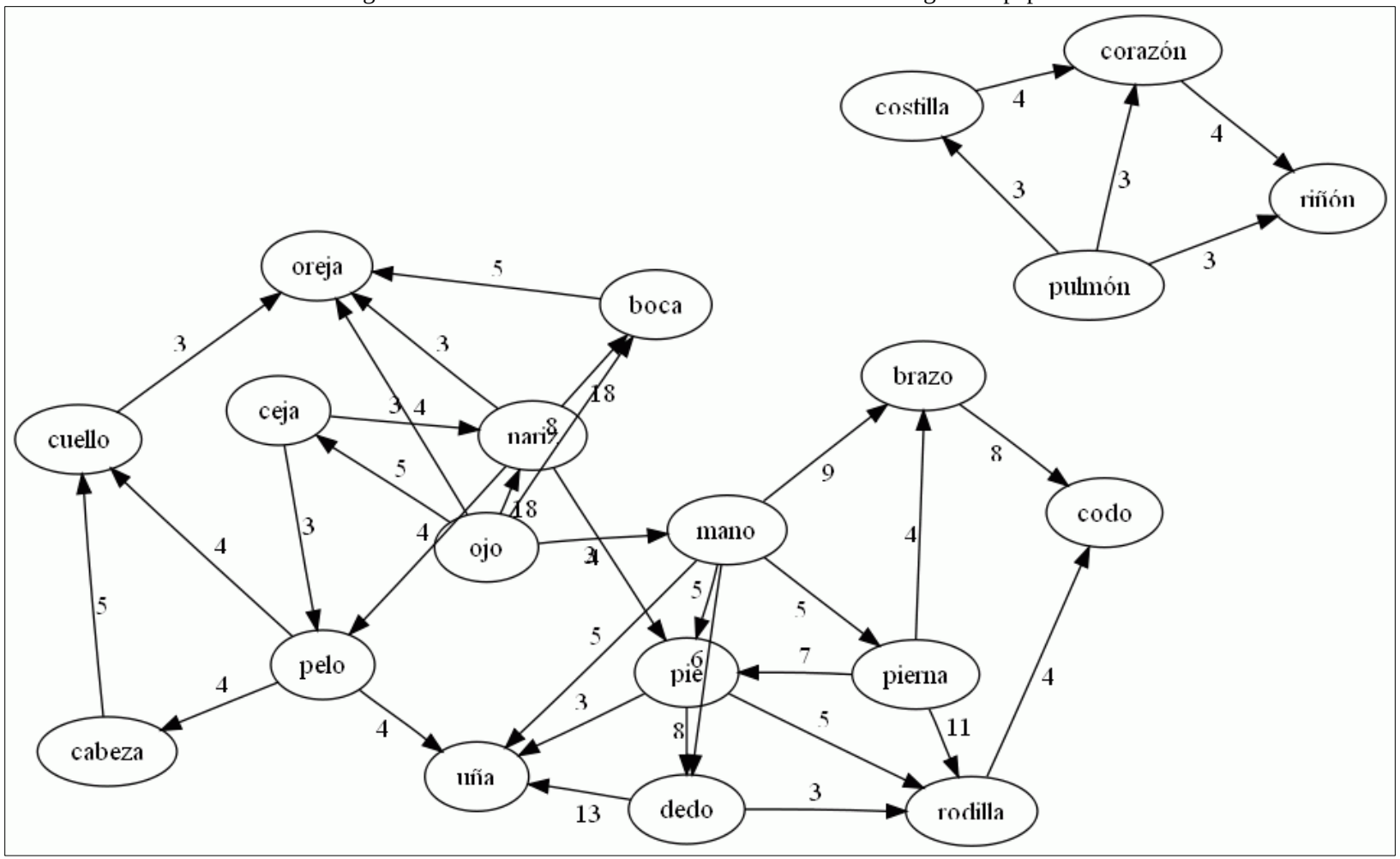
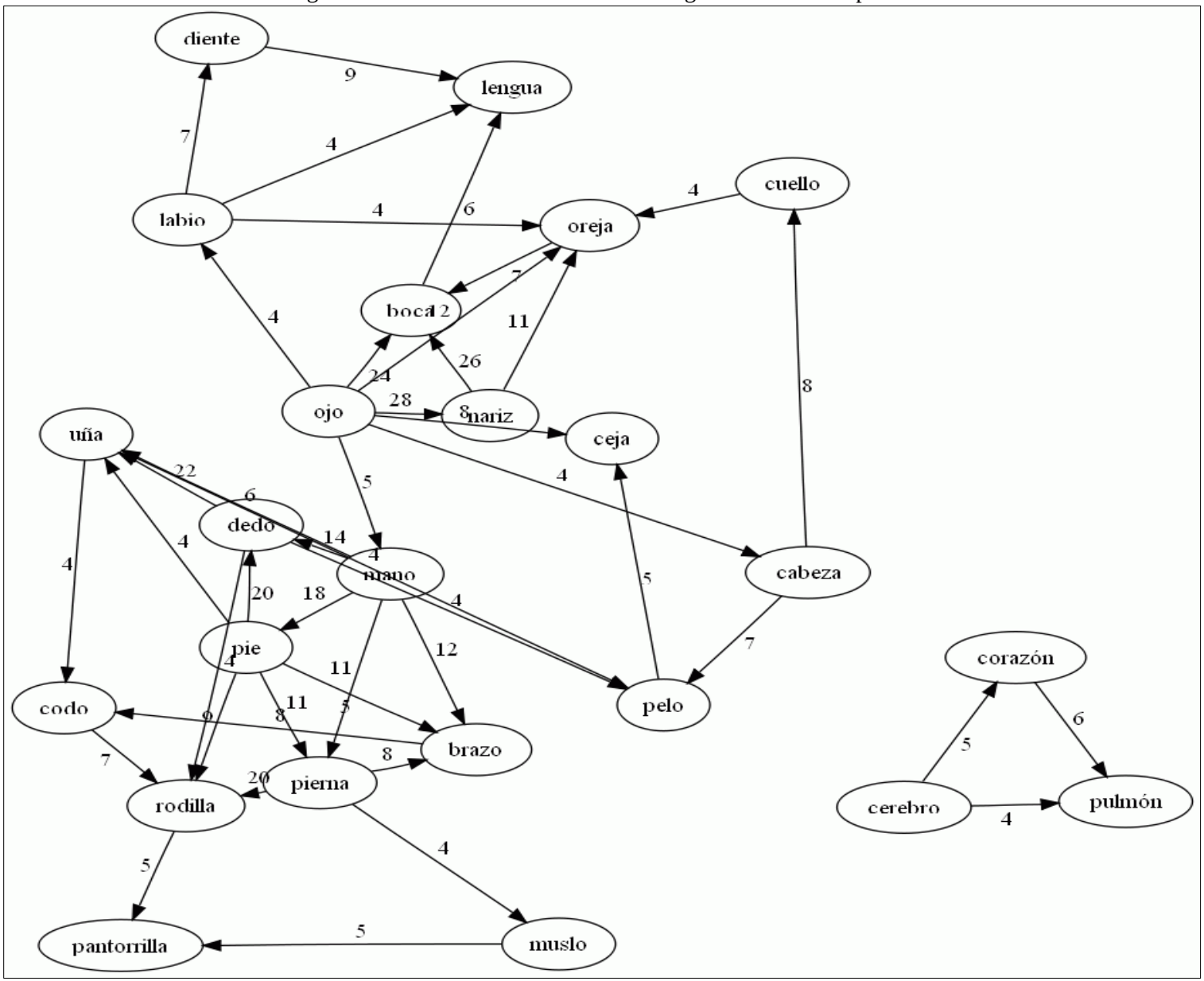


Figura 30. Grafo del CI07 de la muestra digital de Letras Hispánica



En cuanto al conglomerado donde *ojo* es el núcleo, se detalla un mayor número de conexiones, entre las que destacan la afinidad entre *ojo, nariz, boca* y *oreja*, puesto que representan los órganos perceptivos, además, están ubicados en la cabeza, lo que les otorga un carácter de proximidad. Pero también hay que destacar los nexos por holonimia que guardan *ojo-ceja* y *boca-labio*. Por último, *mano* es el centro de un grupo en el que se evidencia afinidades por meronimia: *mano-dedo-uña, mano-brazo, pierna-pie*; tambien, por afinidad *mano-pie*. De este conjunto, resulta interesante la relación *mano-cara*, la cual podría deducirse del hecho de que las personas se pasan las manos por la cara.

En resumen, el análisis de grafos del centro de interés *partes del cuerpo* de las tres muestra bajo estudio corrobora que esta es una categoría compacta y cerrada, en la que las asociaciones semánticas y cognitivas guardan patrones generales bastante similares.

Capítulo 6. Conclusiones

Esta tesis doctoral ha tenido un triple propósito, todos encaminados hacia un mismo fin: iniciar líneas de investigación acerca de algunas características del terreno de la disponibilidad léxica que no se habían procurado o, en su defecto, no se habían tratado a profundidad en estudios previos. En primer lugar, se analizó, cuantitativa y cualitativamente, el léxico disponible de estudiantes universitarios de dos carreras, a saber: Educación Básica y Letras Hispánicas, ambas de la Pontificia Universidad Católica de Chile. La primera se encuentra enumerada en el campo de la Educación, mientras que la segunda, en el de Humanidades y Artes (Unesco-UIS, 2011). A pesar de esto, los dos programas convergen, a grandes rasgos, en algunas aristas, por ejemplo: en la didáctica de la lengua, el lenguaje y la lectura. Sin embargo, cada una las enfoca desde diversas perspectivas.

Esta no es la única pesquisa en la que se analiza el caudal léxico de sujetos pertenecientes a comunidades discursivas dispares (cf. Guerra y Gómez, 2003; Blanco *et al.*, 2020; Herranz, 2020, entre otras), empero sí es una en la que se ha explorado por primera vez, según el arqueo bibliográfico, el caudal léxico de un grupo de Letras Hispánicas. Paralelamente, se ha comparado el vocabulario de la carrera humanística con la pedagógica, tomando en cuenta centros de interés, tanto del PPHLD, como novedosos (*La lectura*, *El profesor*, *La educación*, *Habilidades y cualidades docentes*). Debe indicarse que estos últimos se encuentran –en mayor o menor grado– ligados a las mallas curriculares de los cursos de los encuestados o, en su defecto, a la atmósfera intelectual en la que se circunscriben los participantes. Asimismo, el número de informantes ha superado al de investigaciones en las que también se ha indagado el LD de universitarios, por ejemplo: Guerra y Gómez (2003), Valenzuela *et al.* (2018), Quintanilla y Salcedo (2019), Blanco *et al.* (2020), Fregoso-Peralta y Aguilar-González (2022), Martínez-Lara (2021) y Zambrano (2021).

En segundo lugar, y quizá el aspecto más innovador de esta tesis, se buscaba evaluar un método alternativo al que consuetudinariamente se ha aplicado en los estudios de disponibilidad léxica –sobre todo, desde el marco metodológico del Proyecto Panhispánico de Léxico Disponible (Dispolex, 2023), cuyos fundamentos se desprenden de la propuesta original de Gougenheim *et al.* (1964)–, con el fin de determinar la idoneidad de analizar datos recolectados a través de técnicas vanguardistas, como la aquí postulada. En concreto, el método examinado se sustenta en el empleo de un prototipo digital *ad hoc* gracias al cual se testearon los lexicones, razón por la cual los voluntarios respondieron las encuestas de manera remota u *online*. Aun siendo una técnica progresista, los instrumentos se diseñaron siguiendo *grosso modo* el modelo del PPHLD. Pues, se procuraba mantener los sustentos epistemológicos de la DL y, con ello, disponer de datos comparables. A fin de cumplir con este

objetivo, se contrastaron los vocabularios potenciales de las dos muestras elaboradas por los estudiantes de Letras Hispánicas, ya que –aunque ambos grupos contestaron los test en formatos distintos– los informantes compartían más rasgos extralingüísticos, materiales e inmateriales (cf. Escandell, 2019). Además, los datos eran cuantitativamente adecuados.

Por último, estrechamente vinculado con las dos anteriores metas descritas, esta investigación pretendía dejar, a disposición de la comunidad científica interesada, los léxicos disponibles de los universitarios de Educación Básica y Letras Hispánicas de la PUC, respecto a los centros de interés: *La lectura*, *El profesor*, *La educación*, *Juegos y distracciones*, *La escuela: muebles y materiales*, *Habilidades y cualidades docentes*, *Partes del cuerpo* y *Comidas y bebidas*. A fuer de que se examinaron tres muestras, según la carrera y el formato de las encuestas, el diccionario estadístico creado está organizado en tres partes, a saber: en la primera, se enumeran las palabras disponibles del alumnado de Educación Básica; mientras que la segunda parte contiene el vocabulario de los alumnos de Letras Hispánicas; estos dos productos corresponden a los datos tomados en papel. Y, finalmente, la tercera parte compete al léxico de los humanistas que escribieron los listados digitalmente, mediante a plataforma *ad hoc*.

En atención a estos tres grandes propósitos investigativos, se logró determinar los índices generales de disponibilidad léxica: número de palabras (NP), número de vocablos (NV), promedio de palabras (PP o $\overline{X}$) e índice de cohesión (IC) y densidad léxica (DL), del repertorio elaborado por los 264 informantes, 108 de los cuales cursaban la carrera de Educación Básica (EB) y el restante 156, la de Letras Hispánicas (LH). Los datos de estos últimos, a su vez, se bifurcaban en dos subgrupos, los 68 que realizaron los test en papel y los 88 que lo hicieron digitalmente.

Respecto a la productividad léxica, se indagaron el NP y $\overline{X}$, mientras que sobre la riqueza léxica, se examinó el NV. En concreto, el grupo de EB escribió un total de 14 267 palabras; los de LH 10 722 y los de LH (digital) 13 536. Dejando a un lado los CI07, CI08 y CI05 –por ocupar los rangos más altos en las tres muestras–, el actualizador más productivo (después de estos) en el corpus de EB es *La educación* (1702/ 11,93 %), en el de LH1 es *La lectura* (1266/ 11,81 %) y, en el digital, es *Juegos y distracciones* (1563/ 11,55 %). En cuanto al promedio de palabras, los estudiantes de EB alcanzaron 1783,38 lexías por área nocional y 132,10 por informante; los de LH 1340 por CI y 157,68 por encuestado; por último, los del corpus digital tienen un $\overline{X}$ = 1692 por actualizador y 153,82 por participante. Así pues, se aprecia que –referente al número de palabras– el alumnado de EB expone una productividad mayor que los de LH, concordante con el PP total. Sin embargo, los promedios por sujetos indican que los estudiantes de LH, independientemente del formato de las pruebas, superan

por más de 20 puntos a los de EB. Particularmente sobre los ejes temáticos, los que cuentan con rango (R) cuatro en las muestras 1, 2 y 3, son *La educación* ($\overline{X}$ = 15,76), *La lectura* ($\overline{X}$ = 18,62) y *Juegos y distracciones* ($\overline{X}$ =17,76), respectivamente. Se ha querido hacer este cotejo, puesto que la contaste en los análisis es que los CI 07, 08 y 05 se presentan cuantitativamente como los más ricos y productivos.

En cuanto a la riqueza léxica, las diferencias del número de vocablos entre las muestras de EB y LH no son muy altas, puesto que, si bien el índice de piezas léxicas únicas del grupo de pedagogía (NV = 3941) supera a los cómputos de los corpus de LH (en papel, NV = 3641/ digital, NV = 3934), estos valores son bastante cercanos. Pero, las similitudes son más tangibles entre la muestra 1 y la 3, cuya separación es de apenas 7 vocablos. Contrariamente, este número sube exponencialmente cuando se comparan las dos muestras tomadas en papel, cuya distinción alcanza 300 lemas de separación. El panorama por eje temático exhibe un patrón casi concordante entre las tres muestras, pues el CI *La educación* es el más rico, en las muestras 1 y 2, donde ocupa el R = 1; mientras que en la muestra 3, se halla en el R = 2. Sobre las distinciones destacables, se observa que el actualizador *El profesor* (548/ 13,91 %) se ubica en el R = 2 del corpus de EB. Por su parte, el área temática *Juegos y distracciones* se localiza en el R = 1 (650/ 16,52 %) en los listados de LH tomados digitalmente; en cambio, en el corpus de los discentes del mismo programa académico recolectado físicamente, el CI04 se posiciona en el R = 2 (548/ 15,05 %).

En relación con los análisis de compactibilidad a través del índice de cohesión, estos arrojaron que los CI7, CI8 y CI5 son los más homogéneos o cohesionados. Es decir, los participantes escribieron palabras coincidentes, lo que evidencia que la mayoría de ellos comparten y conocen parcelas léxico-semánticas. Además de estos tres actualizadores, deben resaltarse los que seguían inmediatamente, según los rangos; puesto que las distinciones entre ellos reflejarían, de una u otra manera, aspectos característicos vinculados a los programas de estudio, por ejemplo. Concretamente, en el corpus de EB, los ejes temáticos ubicados en los rangos 4 y 5 son *La lectura* y *La educación*, respectivamente. En tanto que en las dos muestras de LH las áreas nocionales *El profesor* y *La lectura* se ubican en los rangos 4 y 5, ordenadamente.

En consecuencia, se evidencia que –a manera global, independiente de las técnicas y formatos de las pruebas utilizadas– el alumnado de ambos programas presenta pocas diferencias cuantitativas relacionadas con la productividad, la riqueza y la cohesión léxica. Contrariamente, comparten en gran medida ciertos patrones léxicos. Estas cualidades podrían deberse, quizá, a que los informantes pertenecen a un mismo estrato socioeducativo: la comunidad universitaria. Además, en ese mismo contexto, ellos podrían llegar a converger en algunas aristas conceptuales, especialmente sobre la

lengua, la lectura, la docencia y la investigación. Estos espacios comunes de interacción académica podrían fomentar que los hablantes llegasen a compartir algunas experiencias y conocimientos, los cuales se reflejarían a través del léxico. Además, los resultados de cohesión léxica corroboran que (sacando los tres CI más productivos y compactos) las tres muestras concuerdan en posicionar el CI *La lectura* entre los más cohesionados. Pero también marcan algunas preferencias que podrían estar relacionadas con tópicos de sus currículum o intereses profesionales. A manera de ilustración, cuantitativamente, el actualizador *El profesor* ocupa (en el corpus de LH1) el rango 4, mientras que se halla en el R = 6 del corpus de EB.

Sin embargo, cualitativamente, al comparar las veinte palabras más disponibles, las listas de ambas carreras se diferencian por ocho lexías. En el grupo de LH1, los lemas apuntan a conceptos aparentemente de índole general, ligados más a objetos, acciones y espacios de los institutos (*universidad*, *pizarra*, *prueba*, *estudio*, *estudiante*, *plumón*, *sala de clase* y *nota*). Por el contrario, en la muestra de EB, las lexías parecen direccionarse más hacia conceptos enlazados con la personalidad y objetivos de los profesores (*guía*, *educador*, *paciencia*, *apoyo*, *buena onda*, *maestro*, *ayuda* y *persona*). Entonces, podría señalarse que cualitativamente existe una diferencia entre el vocabulario de los alumnos de Educación y los de Letras.

Otro de los objetivos de esta pesquisa apuntaba a conocer el léxico disponible de los grupos bajo análisis sobre ocho centros de interés. Así pues, en las líneas siguientes se detalla el grueso de estas indagaciones. Sobre el CI01 *La lectura*, los cinco vocablos más disponibles del corpus de EB son: *libro*, *leer*, *letra*, *palabra* y *cuento*; las de LH1: *libro*, *letra*, *leer*, *autor* y *palabra*; las de las listas digitales: *libro*, *letra*, *palabra*, *autor* y *leer*. Todas refieren a aspectos generales de esta actividad cognitiva. Las del CI02 se mencionaron en el párrafo previo. Por su parte, las palabras más disponibles del CI03 *La educación* son, en la muestra de EB: *profesor*, *colegio*, *aprendizaje*, *aprender* y *enseñar*; en la de LH1: *profesor*, *colegio*, *universidad*, *aprender* y *alumno*; y en la de LH2: *profesor*, *conocimiento*, *colegio*, *aprendizaje* y *alumno*. Estas representan grandes aristas del concepto explicitado por el estímulo verba, de las que hay que subrayar las que se dirigen a los procesos cognitivos, bases del sistema pedagógico: enseñanza y aprendizaje. Manteniendo el análisis en el área de la didáctica, los cinco vocablos, que encabezan los diccionarios, del CI05 *La escuela* muestran que los tres corpus comparten los lexemas: *mesa*, *silla*, *pizarra* y *lápiz*, distinguiéndose por *plumón* (EB) y *cuaderno* (LH1 y LH2). Sin embargo, se trata de elementos comunes y, por ende, conocidos por los hablantes. En cuanto al CI06 *Habilidades y cualidades docentes*, los listados concuerdan en las lexías: *paciencia*, *empatía*, *enseñar* y *vocación*, pero se diferencian por las palabras *escuchar* (EB) y

comprensión (LH1 y LH2). En este punto, se aprecia que los sujetos tienen un alto grado de coincidencia léxica, en líneas generales, empero cada grupo logra resalta rasgos divergentes del léxico potencial.

Sobre los ejes temáticos tomados del PPHLD. En primer lugar, el CI04 *Juegos y distracciones* es el que expone mayor variabilidad, esto debido a que existe un mayor número de actividades recreativas, juegos de diversos tipos, pasatiempos, etc. De las cinco lexías más disponibles, los corpus comparten las siguientes: *diversión*, *celular* y *amigo*, mientras que los vocablos distintivos por muestras son: *entretención* y *juego de mesa* (EB); *compu(tador/a)* y *carta* (LH1), y *videojuego* (LH2). Debe señalarse que los dos grupos de Letras, además, coinciden en los vocablos: *leer* y *película*; del que se resalta *leer*, que es una afición conectada a la carrera.

Respecto al CI *Partes del cuerpo*, las lexías: *mano*, *ojo* y *cabeza* concuerdan en los tres corpus. Pero las tres muestras se distinguían por *pierna* y *brazo* (EB); y *nariz* y *pie* (muestras 2 y 3). Es decir, los alumnos del área de pedagogía se inclinaron más hacia las extremidades de la fisionomía humana, mientras que los del campo de humanidades fueron aparentemente más ecléticos. Finalmente, el área nocional *Comidas y bebidas* expone coincidencia en tres de las cinco palabras más disponibles, a saber: *Coca-Cola*, *arroz* y *fideo*; pero los corpus se diferencian por los vocablos: *papa frita* y *pizza*, en los listados de EB; y *agua*, *jugo* y *carne*, en los dos de LH. Estos resultados reflejan *grosso modo* una preferencia por la comida rápida y poco saludable por parte de los alumnos de Educación; opuestamente, los de Letras exhiben palabras ligadas a una dieta más saludable. Debe subrayarse que estos datos no intentan medir una preferencia alimenticia.

Ciertamente, sería encomiable poder comparar todos los vocablos disponibles de los ocho centros de interés, teniendo en cuenta los factores *Carrera* y *Formato de la prueba*, pero sería una labor titánica e inabarcable. No obstante, los análisis contrastivos llevados a cabo a partir de las palabras más disponibles favorecieron la posibilidad de conocer con más detalles las convergencias y divergencias generales existente entre los grupos. En relación con lo anterior, debe resaltarse que se observó –salvando las particularidades de cada muestra y centro de interés– una distinción entre las palabras de los dos corpus de Letras Hispánicas frente al de Educación Básica. Particularmente, los diccionarios de los humanistas confluyeron más, lo que no es una sorpresa. Al contrario, era de esperar que –independientemente del tipo de formato de las encuestas– los participantes de LH reportaran palabras similares, ya que comparten un mayor conjunto de conocimientos y experiencias. En este sentido, se evidencia la efectividad en el uso de aplicaciones digitales para la recogida de léxico disponible.

Igualmente, en relación con los propósitos de esta tesis, los análisis cuantitativos, descriptivos e inferenciales, sobre los vocabularios disponibles en relación con los descriptores *Sexo*, *Carrera*, *Año de curso*, *Formato de pruebas*, *Cantidad de libro leídos* y *Frecuencia de lectura* llevados a cabo a través de los cálculos de t de Student y Anova, por medio de SPSS, arrojaron que:

La variable *Sexo*, en la que se empleó el t-test, existe una influencia del grupo femenino en el vocabulario referido a *partes del cuerpo* y *la escuela: muebles y materiales*, en la muestra 1. De forma más acuciosa, el LD del CI07 de las futuras pedagogas se distinguió del de los hombres por los siguientes cuatro lexemas: *rodilla*, *espalda*, *codo* y *estómago*; mientras que los chicos reportaron: *pene*, *pulmón*, *hígado* y *lengua*.

En cuanto al factor *Carrera*, debe acotarse que para este análisis se tomaron en cuenta únicamente los datos recogidos en papel. La prueba paramétrica t de Student indicó que existe una relación estadística significativa entre los programas de estudios y los CI01, CI05 y CI08. También, según el test no paramétrico U de Mann-Whitney, hay una asociación estadística de la variable respecto al número total de palabras y las áreas nocionales CI02, CI04 y CI07, cuyas diferencias léxico-métricas se describieron arriba.

Acerca del factor *Año de curso*, los cálculos realizados mediante el t-test expuso que solo existen incidencias significativas entre los estudiantes del primer año en los ejes temáticos *El profesor*, *La escuela: muebles y materiales*, del grupo de Educación Básica; y en *Comidas y bebidas*, de la muestra digital. Solícitamente, se acota que hay una alta variabilidad en el diccionario de *El profesor*, donde los dos grupos muestran vocablos diversos, como: *enseñar*, *educador*, *buena onda*, *aprender*, *autoridad*, *sabiduría* e *inteligente*, concerniente a los listados del 1.er nivel; y *mediador*, *responsabilidad*, *profesional*, *pedagogía*, *ayuda*, *docencia* y *desafío*, en los lexicones de los del 4.º nivel.

La indagación referida a la variable *Formato de la prueba*, en la que se excluyó el corpus de Educación Básica, evidenció que los datos recolectados a través de internet exhibían medias aritméticas bastante parecidas a la de los vocabularios obtenidos mediante el método tradicional, por lo que las correlaciones por centro de interés no resultaron significativas, salvo en el CI6 *Habilidades y cualidades docentes*. En este se aprecia una incidencia en el caudal léxico de las listas de palabras tomadas en papel, cuyas piezas léxicas exclusivas más disponible son: *pedagogía*, *voz*, *disposición*, *estudio*, *hablar* e *interés*. En cambio, las registradas en el soporte digital son: *responsabilidad*, *escuchar*, *dedicación*, *tolerancia*, *amor* y *perseverancia*. En líneas generales, los resultados cuantitativos y cualitativos han demostrado la confiabilidad del léxico disponible recogido a través de

la plataforma web diseñada y elaborada para esta investigación. Enlazado con este punto previo, Alvar Ezquerra (2004: 20) afirma que "Los corpus están ligados a las nuevas tecnologías y es de esperar que en un futuro inmediato nos ofrezcan unas posibilidades de consulta que se adecúen a nuestras necesidades". En atención a esta proclama, los soportes electrónicos donde se resguardan los corpus son altamente relevantes, por lo que no cabe duda de que en la era digital también se hacen necesarios los instrumentos computacionales que favorezcan la recogida de los materiales que conformarán los corpus.

Para explorar el factor *Cantidad de libros leídos*, primero se recodificó y pasó la variable de cuatro a tres variantes; luego, se aplicó el test paramétrico Anova. Este arrojó que ninguno de los niveles analíticos de la variable condicionó la riqueza léxica de los participantes. Esto pudo deberse a la baja variabilidad entre las variantes, las cuales exponían medias similares en cada muestra, como se observa en el gráfico 36 del subapartado 4.3.

Finalmente, las operaciones realizadas con Anova sobre la variable *Frecuencia de lectura optativa* dieron por significativa la incidencia de este factor en los actualizadores *La lectura*, *La educación*, *Juegos y distracciones* y *Habilidades y cualidades docentes*, únicamente en el corpus digital. No obstante, debe detallarse que los promedios de palabras más altos y, por ende, los influyentes se localizan en la variante 2 (De 1 a 5 horas semanales), para los CI01, CI03 y CI04; por su parte, el mayor PP del CI06 se halla en la variante 3 (De 6 a 10 horas). A pesar de esto, cualitativamente, en el caso del actualizador *La lectura*, el nivel 4 (Más de 10 horas) posee un cómputo superior (8) de vocablos exclusivos y variados, concretamente: *entretención*, *saga*, *librería*, *paper*, *educación*, *texto crítico* e *IVA*. Esto mismo ocurre en el CI04, que detenta 11 lexías únicas: *recreación*, *relajación*, *competencia*, *entretenimiento*, *aire libre*, *disfrute*, *dominó*, *salud (mental)*, *estrés*, *Play(Station)* y *felicidad*. Por último, el CI06 dispone de 12: *transmitir conocimiento*, *aprendizaje*, *moderación*, *enseñanza*, *educación*, *firmeza*, *herramienta pedagógica*, *entendimiento*, *autoridad*, *conciencia*, *crítico* y *evaluación*. Así pues, puede aseverarse que el grado más alto de lectura semanal ostenta la mayor variabilidad que los grados inferiores, aunque estadísticamente no sea significativo.

El último objetivo de esta tesis doctoral se enfoca en los análisis léxicos mediante grafos, que son representaciones gráficas de las asociaciones –a través de mecanismos lógicos y analógicos– que establecen las palabras disponibles (Echeverría *et al.*, 2008; Mateus-Ferro *et al.*, 2018). En este trabajo se optó por explorar las categorías en las que se aglutinan los nodos más relevantes del área nocional *Partes del cuerpo*, a fuer de que es la más productiva, rica y compacta. El examen de los grafos de las tres muestras revela que los nodos se agrupan en dos grandes categorías que aluden, la primera, a la

parte superior del cuerpo: *cabeza/cara*, y la segunda, a las extremidades (superiores e inferiores). Sin embargo, en las dos muestras de Letras Hispánicas se aprecia un tercer conjunto de nodos, separado del resto, que apunta a los órganos internos. En cuanto a los tipos de procesos, las conexiones de las tres muestras se dan máxime por mecanismos lógicos semánticos, particularmente por holónimos y merónimos. Asimismo, se dan casos de encadenamientos por mecanismos analógicos de proximidad espacial/visual y de asociación funcional.

Debe recalcarse que las redes trazadas por Dispografo, a partir de los datos tomados en papel y digitalmente, ofrecen patrones bastante cercanos, tanto en el número de nodos como en la complejidad de los entramados, por lo que puede aseverarse nuevamente que los instrumentos digitales de recolección de léxico disponible pueden llegar a ser altamente confiables.

A manera de cierre, puede señalarse que la metodología desarrollada en este estudio –que integró técnicas novedosas al diseño común del PPHLD, originado de la propuesta de Gougenheim *et al.* (1954, 1964) –, a pesar de sus limitaciones, ha resultado efectiva. En primer lugar, vale recalcar que, por razones externas a la investigación (estallido social y pandemia), se tuvo que adecuar el proyecto original de manera que no se alejara sustancialmente de los primeros objetivos y línea de investigación. En este panorama, la labor más empinada, pero, a la vez, más reconfortante fue la creación de un corpus de DL. Al respecto, debe subrayarse que la construcción de los corpus no es una tarea sencilla, mucho menos cuando el contexto externo a la investigación la complica (Castillo Fadić, 2020; 2021b). El corpus estuvo integrado por materiales recogidos a través de instrumentos de dos formatos: papel y digital, por lo que *grosso modo* eran divergentes y, por ende, representaban una dificultad para llevar a cabo los análisis. No obstante, al contar con dos de las tres muestras bastante cercanas metodológicamente, pudieron llevarse a cabo análisis comparativos que han permitido exponer las ventajas de utilizar aplicaciones electrónicas. Para lograr solventar estas dificultades, fue una fortaleza contar con una tradición disciplinar que ofrece luces sobre los procesos de construcción de corpus, edición, codificación y procesamiento de los materiales.

Segundo, las barreras analíticas pudieron superarse mediante los programas informáticos –especialmente generados para los análisis cuantitativos y cualitativos del LD (Dispogen y Dispografo) y otro auxiliar como IBM SPSS®– para los cálculos de los índices y la producción de los grafos, los cuales llegan a acercarse lo más posible a la realidad lingüística de la población estudiada. No obstante, debe mencionarse que los resultados de esta tesis pudieron robustecerse aún más si se hubieran aplicado análisis estadísticos multivariantes, de manera que se hubiesen determinado los descriptores que, en conjunto, inciden en el caudal léxico de los encuestados. Por lo anterior, queda

abierta la invitación a ahondar en los análisis de los datos. Pero no únicamente en el plano cuantitativo, sino también en el cualitativos, puesto que los diccionarios exhiben ejemplos dignos de exploraciones más detalladas, como: el estudio de las unidades plurilexicales, extranjerismos, chilenismos, tecnicismos, agrupaciones por subcategorías (autores, títulos, sentimientos, valores positivos/negativos), entre otros problemas que no se lograron abarcar.

A pesar de lo anterior, la metodología general puesta en marcha en este estudio permitió, a grandes rasgos, conocer los atributos, cuantitativos y cualitativos, del léxico disponible de estudiantes universitarios de los programas de Educación Básica y Letras Hispánicas de la Pontificia Universidad Católica de Chile, sobre los centros de interés: *La lectura*, *El profesor*, *La educación*, *Juegos y distracciones*, *La escuela: muebles y materiales*, *Habilidades y cualidades docentes*, *Partes del cuerpo* y *Comidas y bebidas*, con lo que se pretendía determinar las variables que pudieran incidir en el caudal léxico de los grupos explorados, así como evaluar la técnica alternativa de recolección de palabras. Con esta panorámica, puede afirmarse que esta tesis doctoral –a pesar de que el corpus fue recogido a través de dos formatos diferentes y con las respuestas de alumnos de dos carreras disímiles– intenta ser un pequeño aporte al conocimiento del vocabulario potencial de los miembros de las comunidades discursivas de Letras Hispánicas y Educación Básica. Además, pretende contribuir con las reflexiones acerca de la ampliación de las técnicas metodológicas de toma y procesamiento del léxico. Sin grandes pretensiones, este trabajo ha querido colaborar con las líneas de investigación realizadas desde la disponibilidad léxica en relación con la sociolingüística.

Aunado a lo anterior, si bien los trabajos léxico-estadísticos suelen enfocarse desde la lingüística aplicada a la enseñanza de lengua, materna o extranjera, esta tesis no buscaba abordar el léxico disponible de los universitarios de Letras y Educación, de la PUC, con pretensiones exclusivamente didácticas. Por el contrario, se centraba más en la descripción y explicación de la conformación de los lexicones, tanto cuantitativa como cualitativamente, con énfasis en las características socioeducativa de los participantes. Así pues, se intenta dejar datos que puedan promover estudios léxicos comparativos de los (sub)grupos que componen las comunidades discursivas universitarias, desde la sociolingüística como a partir de otras disciplinas. No obstante, considerando la afirmación de Ezquerra (1974), quien sostenía que es casi imposible imaginar la elaboración de materiales y programas de enseñanza de cualquier lengua sin antes consultar los diccionarios estadísticos, no cabe duda de que, indirectamente, este trabajo podría ayudar a abrir una pequeña ventana para la examinación de los vocabularios de los discentes de las áreas analizadas. Esto último, además, podría llevar a pensar sobre la planificación lingüística de las carreras.

A partir de las reflexiones de este trabajo, podrían abrirse algunos senderos nuevos en el campo de la disponibilidad léxica o fortalecer algunas líneas de investigación ya existentes. Concretamente, puede seguir ahondándose acerca de las ventajas y limitaciones de los formatos electrónicos para la recolección de listas de palabras potenciales. La bibliografía ya muestra algunos prototipos y experimentos, pero salvo la metodología aquí expuesta, aún quedan por desarrollar trabajos en los que se presenten aplicaciones digitales afinadas, a partir del modelo aquí planteado, que se direccionen a estudios contrastivos más amplios. Particularmente, el acceso a la página web de esta pesquisa es restringido y su ingreso puede ser algo engorroso, por lo que debe avanzarse hacia la afinación de esta para convertirla en una plataforma de acceso libre y más amigable, en la que se entre y se contesten las encuestas de manera expedita, pero controlada. Igualmente, contar con una aplicación web capaz de recolectar efectivamente el léxico disponible de un grupo puede motivar la realización de experimentos en los que se calcule, no solo los índices de DL, sino también el tiempo de respuesta y escritura de palabras en el teclado. Y, por supuesto, a medida que más participantes intervengan en los test digitales, podría ir armándose una base de datos mayor.

Por último, deben impulsarse las investigaciones contrastivas del léxico disponible de comunidades discursivas diferentes, con las que pueda trazarse la composición de los lexicones de las distintas áreas del conocimiento de los programas académicos, sobre todo en una época marcada por la interdisciplinariedad. En esta misma línea, se contribuye con el conocimiento y reporte del vocabulario potencial de áreas nocionales nada o poco descritas. Si bien el léxico disponible aquí analizado ha dado luces respecto a campos nocionales nuevos o tratados en otras sintopías, en futuras pesquisas pueden analizarse las carreras pertenecientes a una misma facultad, con la finalidad no solo de determinar el vocabulario utilizado potencialmente por los universitarios, sino también con el objetivo de evaluar la adquisición y estabilidad de los conocimientos especializados. En este punto, cabría la posibilidad de profundizar sobre los estudios terminológicos a partir de los corpus de disponibilidad léxica. En el caso de Letras PUC, específicamente, podría analizarse el LD referido, no solo a la lectura, sino también a ciencias y disciplinas, como: lingüística, literatura, teoría literaria, entre otros. Además, podría abordarse el vocabulario en español (Letras Hispánicas) e inglés (Letras Inglesas). Como se ve, esta tesis ha sido apenas un ápices en la línea de investigación de la disponibilidad léxica.

Referencias bibliográficas

AABIDI, Lahoussine, 2020: "La disponibilidad léxica en español de alumnos marroquíes de enseñanza media: resultados generales", *Philologica Canariensia* 26, 1-19.

AITCHISON, Jean. 1993: *El cambio en las lenguas: ¿progreso o decadencia?* Barcelona: Ariel.

ALAMEDA, José, y Fernando CUETOS, 1995: *Diccionario de frecuencias de las unidades lingüísticas del castellano,* Vol. 1 y Vol. 2, Oviedo: Servicio de Publicaciones de la Universidad de Oviedo.

ALBA, Orlando, 1995: *Léxico disponible de la República Dominicana,* Santo Domingo: Pontificia Universidad Católica "Madre y Maestra".

ALMELA PÉREZ, Ramón. 1999: *Procedimientos de formación de palabras en español*. Barcelona: Ariel.

ALVAR, Manuel. 1996: *La formación de palabras en español*. Barcelona: Ariel.

ALVAR EZQUERRA, Manuel, 2004: "La frecuencia léxica y su utilidad en la enseñanza del español como lengua extranjera", *Actas del XV Congreso de la ASELE,* 19-39.

ÁLVAREZ-ÁLVAREZ, Carmen, y José Manuel DIEGO-MANTECÓN, 2019: "¿Cómo describen, analizan y valoran los futuros maestros su formación lectora?", *Revista Complutense de Educación* 30 (4), 1083-1096.

ATKINS, B. T. Sue y Michael RUNDELL: 2008. *The Oxford Guide of Practical Lexicography*. Oxford: Oxford University Press.

ARAVENA, Soledad y HUGO, Evelin. 2017: "Desarrollo de la complejidad sintáctica en textos narrativos y explicativos escritos por estudiantes secundarios". *Lenguas Modernas*, 47, 9-40 [Recuperado: https://enfoqueseducacionales.uchile.cl/index.php/LM/article/view/45181, fecha de consulta: 20 de julio de 2023]

ARAVENA, Soledad y QUIROGA, Riva. 2018: "Desarrollo de la complejidad léxica en dos géneros escritos por estudiantes de distintos grupos socioeconómicos", *Onomázein*, 42, 197-224 [Recuperado: https://doi.org/10.7764/onomazein.42.03, fecha de consulta: 20 de julio de 2023]

ARNAL, Mª Luisa, 2008: "Niveles socioculturales y léxico dialectal en el vocabulario disponible de Aragón", en Blas, J. L. *et al*. (ed), *Discurso y sociedad II. Nuevas contribuciones al estudio de la lengua en contexto*. 457-569. Universidad Jaume I.

ARNAL, Mª Luisa, Rosa CASTAÑER, José María ENGUITA, Vicente LAGÜÉNS y Ana MOLINÉ, 2004: *Léxico disponible de Aragón*, Zaragoza: Libros Pórtico.

ARISTIZÁBAL, Enrique, 1938: *Détermination Expérimentale du Vocabulaire Écrit pour servir de Base à l'Énseignement de l'Orthographie a l'École Primaire*, París: H. Champion.

AUBER, M, 1953: *Guide de fréquence*, París: Gap Ophrys.

ÁVILA MARTÍN, M.ª del Carmen, 2010: "Estadística y lingüística de corpus: implicaciones pedagógicas en la enseñanza y el aprendizaje del léxico" *Cauce. Revista internacional de Filología, Comunicación y sus Didácticas* 32, 163-175.

ÁVILA MUÑOZ, Antonio, 2006: *Léxico disponible de los estudiantes preuniversitarios malagueños*. Málaga: Servicio de Publicaciones de la Universidad de Málaga.

ÁVILA MUÑOZ, Antonio, 2007: "Léxico disponible y ortografía. condicionantes sociales y hábitos culturales de influencia". En Juan Moya y Marcin Sosinsky (eds.). *Las hablas andaluzas y la*

enseñanza de la lengua. Actas de las XII Jornadas sobre la Enseñanza de la Lengua Española, pp. 25-46. Granada: Universidad de Granada.

ÁVILA MUÑOZ, Antonio, 2010: "Introducción. Aproximación a los estudios estadísticos del léxico. Teoría y principio". En Antonio ÁVILA MUÑOZ y Juan VILLENA (ed.): *Variación social del léxico disponible en la ciudad de Málaga. Diccionario y análisis*, pp. 15-34. Málaga: Editorial Sarriá.

ÁVILA MUÑOZ, Antonio, 2016: "El léxico disponible y la enseñanza del español. Propuesta de selección léxica basada en la teoría de los conjuntos difusos". *Journal of Spanish Language Teaching* (3)1, 31-43.

ÁVILA MUÑOZ, Antonio y José SÁNCHEZ, 2011: "La posición de los vocablos en el cálculo del índice de disponibilidad léxica: Proceso de reentrada en las listas del léxico disponible en la ciudad de Málaga", *ELUA* 25, 45-74.

ÁVILA MUÑOZ, Antonio, y José SÁNCHEZ-SÁEZ, 2014: "Fuzzy sets and prototype theory. Representational modelo f cognitive community structures based on lexical availability trials", *Review of Cognitive Linguístics* 12:1, 133-159.

ÁVILA MUÑOZ, Antonio y María LASARTE, 2010: "Estudio del conocimiento social del léxico disponible. Objetivos, metodología y criterios de edición del diccionario". En Antonio ÁVILA MUÑOZ y Juan VILLENA (ed.): *Variación social del léxico disponible en la ciudad de Málaga. Diccionario y análisis*, pp. 83-112. Málaga: Editorial Sarriá.

ÁVILA MUÑOZ, Antonio, José SÁNCHEZ-SÁEZ y Nana ODISHELIDZE, 2021: "DispoCen. Mucho más que un programa para el cálculo de la disponibilidad léxica", *ELUA* 35, 9-36.

ÁVILA, Antonio; SANTOS-DÍAZ, Inmaculada y TRIGO IBÁÑEZ, Ester, 2020: "Análisis léxico-cognitivo de la influencia de los medios de comunicación en las percepciones de universitarios españoles ante la COVID-19", *Círculo de Lingüística Aplicada a la Comunicación*, 84, 85-95.

ÁVILA MUÑOZ, Antonio y Juan VILLENA (eds.), 2010: *Variación social del léxico disponible en la ciudad de Málaga. Diccionario y análisis*. Málaga: Editorial Sarriá.

AYRES, L, 1915: *A Measuring Scale for Ability in Spelling*, New York: The Russell Sage Foundation.

AYORA, C., 2004: *Disponibilidad léxica en Ceuta: aspectos sociolingüísticos*, Cádiz: Universidadde Cádiz.

AZURMENDI, María, 1982: *Elaboración de un modelo para descripción sociolingüística del bilingüismo y su aplicación parcial en la comarca de San Sebastián*. Tesis Doctoral. Madrid: Universidad Complutense de Madrid.

BASIC-ENGLISH INSTITUTE. 2001-2014: "Ogden's Basic English" [Recuperado: http://ogden.basic-english.org/basiceng.html, fecha de consulta: 20 de julio de 2023]

BAILEY, J, 1971: *A Study of Lexical Availability among Monolingual/Bilingual Speakers of Spanish and English*. Tesis de maestría. Houston: Rice University.

BARRIO, Florencio del, 2016: "Algunas observaciones sobre la disponibilidad léxica en estudiantes italófonos de español", en Eugenia SAINZ, Inmaculada SOLÍS, Florencio del BARRIO, Ignacio ARROYO (eds.), *Geométrica explosión: Estudios de lengua y literatura en homenaje a René Lenarduzzi*, Venezia: Università Ca' Foscari Venezia, 127-144.

BARTOL, José, 2006: "La disponibilidad léxica", *Revista Española de Lingüística (RSEL)*, 379-396.

BARTOL, José y Natividad HERNÁNDEZ, 2005: "Proyecto del léxico disponible de España. Discurso y Sociedad. Contribuciones al estudio de la lengua en contexto social", en José Luis BLAS ARROYO,

Manuela CASANOVA y Mónica VELANDO, Castellón de La Plana: Servicio de Publicaciones de la Universidad Jaume I, 731-742.

BENÍTEZ, Pedro, 1995: "Disponibilidad léxica en Madrid: Análisis cuantitativo", en Actas del III Seminario Internacional sobre "Aportes de la Lingüística a la enseñanza de la lengua materna", *Revista de la Adquisición de la Lengua Española*, número especial.

BENÍTEZ, Pedro, 2003: "Consideraciones en torno a la enseñanza del vocabulario" en Francisco MORENO FERNÁNDEZ, Francisco GIMENO, José A. SAMPER, María Luz GUTIÉRREZ, María VAQUERO y César HERNÁNDEZ (coords.), *Lengua, Variación y contexto*, vol. I, Madrid: Arco/Libros, 145-165.

BENÍTEZ, Pedro, Clara Eugenia HERNÁNDEZ y José A. SAMPER PADILLA, 1995: "Léxicos básicos de España (LEBADES) y de Canarias (LEBAICan). Proyecto de Investigación", *REALE*, 3, 9-17.

BRIONES, Guillermo, 1990: *Métodos y técnicas de investigación para las ciencias sociales*, México: Trillas.

BLANCA, María, Rafael ALARCÓN, Jaume ARNAU, Roser BONO y Rebecca BENDAYAN, 2017: "Non-normal data: Is ANOVA still a valid option?", *Psicothema* 29, 4, 552-557.

BLANCO CORREA, Óscar, Pedro SALCEDO y Gabriela KOTZ GRABOLE, 2020: "Análisis del léxico de las emociones: una aproximación desde la disponibilidad léxica y la teoría de los grafos léxicos", *Lingüística y Literatura* 78, 55-83.

BOLÍVAR, Adriana, 2005: "Tradiciones discursivas y construcción del conocimiento en las humanidades", *Signo y Seña* 14, 67-91.

BOLÍVAR, Adriana, 2013: "La definición de los corpus en los estudios del discurso", Revista Latinoamericana de Estudios del Discurso 13, 1, 3-8.

BOSQUE, Ignacio. 2001: "Sobre el concepto de 'colocación' y sus límites", *Lingüística Española Actual* 23, 1, 9-40.

BUCHANAN, Milton, 1929: *A graded Spanish word book*. Toronto: The University of Toronto Press. https://babel.hathitrust.org/cgi/pt?id=mdp.39015019056624&view=1up&seq=5

BUTRÓN, Gloria, 1991: "Nuevos índices de disponibilidad léxica". En López Morales, Humberto (ed.). *La enseñanza del español como lengua maternal*. 79-89. Río Piedar: Editorial de la Universidad de Puerto Rico. https://books.google.es/books?hl=es&lr=&id=7pdU-ht5N-4C&oi=fnd&pg=PA79&dq=Gloria+Butr%C3%B3n+Nuevos+%C3%ADndices+de+disponibilidad+l%C3%A9xica&ots=rzGGy80GIy&sig=X-6uwtLa_tXATJdu8eybNwGv-TI#v=onepage&q=Gloria%20Butr%C3%B3n%20Nuevos%20%C3%ADndices%20de%20disponibilidad%20l%C3%A9xica&f=true

CALLEALTA, Francisco y Diego GALLEGO, 2016: "Medidas de disponibilidad léxica: Comparabilidad y normalización", *Boletín de Filología*, LI, 1, pp. 39-92.

CAMPILLOS, Leonardo, e Hiroto UEDA, 2015: "Frecuencia y dispersión léxica en textos médicos divulgativos en español", *Ibérica* 30, 61-84.

CAÑIZAL ARÉVALO, Alva, 1987: *Disponibilidad léxica en escolares de primaria terminada. Análisis de seis centros de interés*. Tesis de Licenciatura en Lengua y Literatura Hispánica. Universidad Autónoma de México.

CARCEDO, Alberto, 2003: "Unidad y variedad diatópica de la disponibilidad léxica del español: comparación de los inventarios de Puerto Rico, Cádiz y Asturias", en Francisco MORENO FERNÁNDEZ,

Francisco GIMENO, José SAMPER PADILLA, María Luz GUTIÉRREZ, María VAQUERO y César HERNÁNDEZ (coord.) *Lengua, variación y contexto*, vol. I, 199-225. Madrid: Arco/Libros.

CÁRDENAS, Manuel y Héctor ARANCIBIA, 2014: "potencia estadística y cálculo del tamaño del efecto en g*power: complementos a las pruebas de significación estadística y su aplicación en psicología", *Salud & Sociedad* 5, 2, 210-224.

CARRERO FERNÁNDEZ (ed.), *Aplicaciones de la disponibilidad léxica*, Universitat de Lleida, Lleida, España.

CASANOVA, Manuela, 2017: *Léxico disponible de Castellón. Estudio y diccionarios*. Tesis para optar al grado de doctora, Universitat Jaume I.

CASTILLO FADIĆ, M.ª Natalia, 2021a: *Léxico básico del español de Chile*, Santiago: Liberalia Ediciones.

CASTILLO FADIĆ, M.ª Natalia, 2021b: "Corpus Básico del Español de Chile ©: metodología de obtención, revisión y constitución definitiva". En Abelardo SAN MARTÍN, Darío ROJAS y Soledad CHÁVEZ (editores), *Anejo N°3 del Boletín de Filología, Estudios en homenaje a Alfredo Matus Olivier*, 219-252.

CASTILLO FADIĆ, M.ª Natalia y Josué PINO CASTILLO, 2020: "Hacia la construcción de un instrumento para evaluar la familiaridad de pacientes crónicos con unidades léxicas relevantes para el automanejo de su condición de salud", *Nueva Revista del Pacífico* 72, 86-115.

CASTILLO FADIĆ, M.ª Natalia e Inmaculada, SANTOS DÍAZ, 2021: "Dos miradas sobre el cuerpo humano en el marco de la literacidad en salud: Profesionales de la salud y personas con diabetes Mellitus 2", en Maribel Serrano y Mª. Ángeles Calero (Eds.), *Aplicaciones de la disponibilidad léxica*. Sevilla: tirant humanidades, 277-293.

CASTILLO FADIĆ, M.ª Natalia y Enrique SOLOGUREN, 2018: "Pretérito imperfecto de subjuntivo en el español de Chile: ¿existe alternancia libre entre las desinencias -ra y -se?" *Onomázein* 42, 153-171.

CEPEDA, Milko, 2017: "Índices de disponibilidad léxica en alumnos de prekínder en un contexto rural de la comuna de Maule, Chile", *Interpretextos* 18, 183-199.

CEPEDA, Milko, Ángela CÁRDENAS, Macarena CARRASCO, Nicole CASTILLO, Joselyne FLORES, Constanza GONZÁLEZ y Melanie ORÓSTICA, 2016: "Relación entre disponibilidad léxica y comprensión lectora, en un contexto de educación técnico profesional rural", *Sophia Austral* 18: 81-93.

CEPEDA, Milko, Maribel GRANADA y María POMÉS, 2014: "Disponibilidad léxica en estudiantes de primero básico", *Literatura y Lingüística* 30, 166-181.

CERDA, Macarena, 2014: *Reconocimiento auditivo de palabras y comprensión oral de textos descriptivos en niños preescolares*. Tesis de maestría, Universidad de Concepción de Chile.

CERDA, Gamal, Pedro SALCEDO, Carlos PÉREZ y Verónica MARÍN: 2017. "Futuros Profesores de Matemáticas: Rol de la Disponibilidad Léxica, Esquemas de Razonamiento Formal en Logros Académicos Durante su Formación Inicial", *Formación universitaria* 10, 1, 33-46. https://dx.doi.org/10.4067/S0718-50062017000100005

CERDA ETCHEPARE, Gamal, Carlos PÉREZ WILSON y Eugenio CHANDÍA MUÑOZ, 2021: "Disponibilidad léxica en centros de interés asociados a ejes curriculares de matemáticas en estudiantes de contextos de alta vulnerabilidad social", *Revista de Estudios y Experiencias en Educación (REXE)* 42, 33-51.

COLOMER, Teresa y Felipe MUNITA, 2013: "La experiencia lectora de los alumnos de Magisterio: nuevos desafíos para la formación docente", *SEDLL. Lenguaje y Textos* 38, 37-44.

CUENCA, María José y Joseph HILFERTY, 1999: *Introducción a la lingüística cognitiva*, Ariel, Barcelona.

DEL BARRIO, Florencio y Félix SAN VICENTE. 2015: "La formación de palabras". En Felix SAN VICENTE (Dir.) *GREIT. Gramática de referencia de español para italófonos*, 1413-1520. Bologna: CLUE y EUS.

DEL MASTRO, F. y SUMAR, O. A. 2021: "Del salón al Zoom: cambios en la interacción docente-estudiante en dos facultades de derecho del Perú", *Revista Pedagogía Universitaria Y Didáctica Del Derecho*, *8*(2), 119-150. https://doi.org/10.5354/0719-5885.2021.61786

DEL VALLE, María, Pedro SALCEDO y Anita FERREIRA, 2016: "Analyzing the Availability of Lexicon in Mathematics Education Using no Traditional Technological Resources", *Journal of Supply Chain Management* Vol. 5, No. 2, 144-149.

DIMITRIJEVIĆ, Naum, 1969: *Lexical availability*. Alemania: Julius Groos Verlag Heidelberg.

DI TULLIO, Ángela. 1998. *Manual de gramática del español*. Buenos Aires: Edical.

DISPOLEX (2023): *¿Qué es el Proyecto Panhispánico?* <http://www.dispolex.com/info/el-proyecto-panhispanico>

DORTA, G. 2008: "Politeness and Social Dynamics in Chat Communication". En Grein, M (Ed.), *Dialogue in and between Different Culture*. Pp. 111-124. iada.online.series

ECHEVERRÍA, Max, 1991: "Crecimiento de la disponibilidad léxica en estudiantes chilenos de nivel básico y medio", en Humberto LÓPEZ MORALES (ed.), "*La enseñanza del español como lengua materna*. Río Piedras: Universidad de Puerto Rico, 61-78.

ECHEVERRÍA, Max, 2001: "Estructura y funciones de un software de vocabulario disponible", *RLA: Revista de lingüística teórica y aplicada* 39, 87-100.

ECHEVERRÍA, Max S., MARÍA OLIVIA Herrera, PATRICIO Moreno y Francisco PRADENAS, 1987: "Disponibilidad léxica en Educación Media", en *Revista de Lingüística Teórica y Aplicada* 25, 55-115.

ECHEVERRÍA, Max y Mabel URRUTIA, 2004: "Incidencia del envejecimiento en el acceso léxico", *Revista Chilena de Fonoaudiología* 5, 2, 7-23.

ECHEVERRÍA, Max S., Paula URZÚA e Israel FIGUEROA, 2005: *Dispogen II. Programa computacional para el análisis de la disponibilidad léxica*. Concepción, Chile: Universidad de Concepción.

ECHEVERRÍA, Max, Roberto VARGAS, Paula URZÚA y Roberto FERREIRA, 2008: "DispoGrafo: Una nueva herramienta computacional para el análisis de relaciones semánticas en el léxico disponible", *Revista de Lingüística Teórica y Aplicada* 46, 81-91.

ELCHE LARRAÑAGA, María, y Santiago YUBERO JIMÉNEZ, 2019: "La compleja relación de los docentes con la lectura: El comportamiento lector d*el profesor*ado de educación infantil y primaria en formación", *Bordón* 71(1), 31-45.

ESCANDELL, M.ª Victoria, 2019: *Introducción a la pragmática*. 7.ª impresión. Barcelona: Ariel/Letras.

ESCUDERO, Rocío, SANTOS DÍAZ, Inmaculada y Ester TRIGO IBÁÑEZ, 2022: "Evolución del léxico disponible sobre la "Escuela" del alumnado de Educación Primaria según el tipo de centro educativo", *Cultura, Lenguaje y Representación,* XXVIII, 61–81.

ESPINAL, María Teresa y Jaume MATEU, 2014: "Palabras y significado", en María Teresa ESPINAL, *Semántica*, Madrid, Akal, 59-110.

EZQUERRA, Raimundo, 1974: "Los diccionarios de frecuencia en español", Boletín de la AEPE 10, 3-27.

FAINHOLC, B. 2021: "El ZOOM y la educación", *Didáctica, Innovación y Multimedia*, 39. https://raco.cat/index.php/DIM/article/view/388779

FASCE, Eduardo, Max ECHEVERRÍA, Olga MATUS, Liliana ORTIZ, Sylvia PALACIOS, Alejandro SOTO, 2009: "Atributos del profesionalismo estimados por estudiantes de Medicina y médicos. Análisis mediante el modelo de disponibilidad léxica", *Revista Médica de Chile* 137, 746-752.

FAUL, Franz, Edgar ERDFELDER, Albert-Georg LANG y Axel BUCHNER (2007): "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences", en *Behavior Research Methods, 39*, pp. 75-191.

FAUL, Franz, Edgar ERDFELDER y Albert-Georg LANG (2009): "Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses", en *Behavior Research Methods, 41*, pp. 1149-1160.

FELÍU, Elena, 2009: "Palabras con estructura interna", en *Panorama de la lexicología*. Elena de MIGUEL (ed.), Barcelona, Ariel, 51-82.

FERNÁNDEZ JUNCAL, M.ª Carmen, 2008: *Léxico disponible de Burgos*. Burgos, Instituto Castellano y Leonés de la Lengua.

FERNÁNDEZ JUNCAL, M.A Carmen y Natividad HERNÁNDEZ, 2021: "Antropónimos disponibles: consolidación de modelos socionomásticos", en Maribel SERRANO y Mª. Ángeles CALERO (Eds.), *Aplicaciones de la disponibilidad léxica*. Sevilla: tirant humanidades, 21-276.

FERREIRA, Anita, Pedro SALCEDO y María DEL VALLE, 2014: "Estudio de disponibilidad léxica en el ámbito de las matemáticas", *Estudios Filológicos* 54, 69-84.

FERREIRA, Roberto y Max ECHEVERRÍA, 2010: "Redes semánticas en el léxico disponible de inglés L1 e inglés LE", *Onomázein* 21, 133-153.

FERREIRA, Roberto, Jaime GARRIDO MOSCOSO y Alexia GUERRA RIVERA, 2019: "Predictors of lexical availability in English as a second language", *Onomázein* 46, 18-34.

FRAGUELA-VALE, Raúl, Héctor POSE-PORTO y Lara VARELA-GARROTE, 2016: "Tiempos escolares y lectura", *Ocnos: Revista de Estudios sobre lectura* 15 (2), 67-76.

FREGOSO-PERALTA, Gilberto, y Luz Eugenia AGUILAR-GONZÁLEZ, 2022: "Disponibilidad, diversidad y complejidad de vocabulario técnico en estudiantes de veterinaria", *Revista de Educación y Desarrollo* 60, 41-51.

FREY, María L, 2007: "Disponibilidad léxica y escritura del español como lengua extranjera: propuesta de comparación de dos corpus", *Interlingüística* 17, 366-373.

FREY, María L, 2008: *El léxico disponible en los escritos de alumnos de español como lengua extranjera. Estudio comparativo de los córpora*. Tesis doctoral, Universidad de Alcalá en Alcalá.

FRIES, Charles, y Aileen TRAVER, 1940: *English word lists; a study of their adaptability for instruction*, Washinton: American Council on Education, University of Michigan. https://babel.hathitrust.org/cgi/pt?id=mdp.39015019387169&view=1up&seq=9 .

FRÖHLICH, A. y S. LUX. 2008: "Chat Behaviour Intercultural. Some aspects of chat behaviour". En Grein, M (Ed.), *Dialogue in and between Different Culture*. Pp. 125-150. iada.online.series

GALLOSO CAMACHO, Mª. Victoria, 1998: *El léxico disponible en el nivel preuniversitario (provincia de Zamora)*, Salamanca: Universidad de Salamanca.

GALLOSO CAMACHO, María, y Montemayor MARTÍN CAMACHO, 2021: "Disponibilidad léxica, diccionario de onubensismos y enseñanza de la lengua", *Revista sobre investigaciones léxicas RILEX* 4, 95-121.

GARCÍA MEGÍA, Antonio, 2003: *La disponibilidad léxica en la ciudad de Almería*. Tesis doctoral inédita, Universidad de Almería en Almería.

GARCÍA-PAGE, Mario. 2008: *Introducción a la fraseología española. Estudio de las locuciones*. Barcelona: Anthropos.

GARCÍA PLATERO, Juan, 2015: "La motivación léxica en el ámbito de la glosodidáctica", *ELUA* 29, 155-170.

GARZÓN, Anni y Luis PENAGOS, 2016: "Disponibilidad léxica en estudiantes de primer semestre de pregrado de una institución universitaria de Villavicencio, Colombia", *Forma y Función* 29, 1, 63-84.

GIBBS, Raymond, 1996: "What's cognitive about cognitive Linguistics?", en CASAD Eugene (ed), *Linguistics in the Redwood: The Expansion of a New Paradigm in Linguistics*. Berlin, Mouton de Gruyter, 27-53.

GÓMEZ DEVÍS, M.ª Begoña, 2004: *La disponibilidad léxica de los estudiantes preuniversitarios valencianos: reflexión metodológica, análisis sociolingüístico y aplicaciones*. Tesis Doctoral, Universitat de Valencia.

GÓMEZ DEVÍS, M.ª Begoña, 2019: "A propósito de las redes semánticas en el léxico disponible de escolares de primero de Educación Primaria", *Ogigia. Revista Electrónica De Estudios Hispánicos* 25, 165-183. https://doi.org/10.24197/ogigia.25.2019.165-183

GÓMEZ DEVÍS, M.ª Begoña, 2021: "Disponibilidad léxica en niños de 6 años. Alcance y proyección didáctica del corpus léxico infantil", *Cultura, Lenguaje y Representación* Vol. XXV, 169-181.

GÓMEZ DEVÍS, M.ª Begoña y Milko CEPEDA, 2022: "Bases para la enseñanza del léxico: mecanismos de asociación y configuración de redes en el léxico disponible infantil", *Tejuelo* 35, 3, 105-134.

GÓMEZ DEVÍS, M.ª Begoña y Cristina HERRANZ, 2022: "Léxico disponible de escolares de la etapa primaria o básica: bases y propuesta metodológicas", *Pragmalingüística* 30, 183-204.

GÓMEZ MOLINA, José Ramón, 2021: "La disponibilidad léxica y sus aplicaciones cognitivas", en Maribel Serrano y Mª. Ángeles Calero (Eds.), *Aplicaciones de la disponibilidad léxica*. Sevilla: tirant humanidades, 207-230.

GÓMEZ MOLINA, José Ramón, y M.ª Begoña GÓMEZ DEVÍS, 2004: *La disponibilidad léxica de los estudiantes preuniversitarios valencianos. Estudio de estratificación sociolingüística*, Valencia: Universitat de Valencia.

GONZÁLEZ, Antonio, Santiago CABANES y Francisco GARCÍA. 1982: *Léxico básico de la lengua escrita en la República Dominicana*. Santo Domingo, Dirección de publicaciones de la Universidad Nacional Pedro Henríquez Ureña.

GONZÁLEZ FERNÁNDEZ, Javier, 2014: "Idoneidad de los centros de interés clásicos en los estudios de disponibilidad léxica aplicados al español como lengua extranjera", *Revista Nebrija de Lingüística Aplicada* 16, (SP).

GORDEJUELA SENOSIÁIN, Adriana, Dámaso IZQUIERDO ALEGRÍA, Felipe JIMÉNEZ BERRIO, Alberto DE LUCAS VICENTE y Manuel CASADO VELARDE (Eds.), 2015: *Lenguas, lenguaje y lingüística. Contribuciones desde la Lingüística General*, Pamplona: Universidad de Navarra.

GOUGENHEIM, Georges, 1954: "Le français élémentaire", *The french review* 27, 3, 217-220.

GOUGENHEIM, Georges, 1955: "Le Français élémentaire. Étude sur une langue de base", *International Review of Education / Internationale Zeitschrift für Erziehungswissenschaft / Revue Internationale de l'Education* 1, 4, 401-412

GOUGENHEIM, Georges, René MICHÉA, Paul RIVENC y Aurélien SAUVAGEOT, 1954: *L'élaboration du français élémentaire*. Paris: Didier.

GOUGENHEIM, George, René MICHÉA, Paul RIVENC y Aurélien SAUVAGEOT, 1964: *L'Élaboration du Français Fondamental (1 degré). Étude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris: Dider.

GRANADO, Cristina, y María PUIG, 2014: "¿Qué leer los futuros maestros y maestras? Un estudio del docente como sujeto lector a través de los títulos de libros que evocan", *Ocnos* 11, 93-112.

GUERRA SALAS, Luis, y M.ª Elena GÓMEZ SÁNCHEZ, 2003: *Español de los medios de comunicación: aspectos de disponibilidad léxica*, comunicación presentada en el XIV Congreso Internacional de ASELE, Burgos.

GUIRAUD, Pierre, 1960: *Problème et méthodes de la statistique linguistique*. Países Bajos: Springer.

GUNTER, Barrie, 2014: "Los procedimientos de las investigaciones cuantitativas". En BRUHN JENSE, Klaus (ed.): *La comunicación y los medios. Metodología de investigación cualitativa y cuantitativa*. México: Fondo de Cultura Económica. 342-384.

HENRÍQUEZ GUARÍN, María Clara, Viviana MAHECHA MAHECHA y Geral Eduardo MATEUS FERRO, 2016: "Análisis de los mecanismos cognitivos del léxico disponible de cuerpo humano a través de grafos", *Lingüística y Literatura* 69, 229-251.

HENMON, V.A.C, 1924: "A French word book based on a count of 400,000 running words", *Bureau of Educational Research Bulletin* 3, Madison: University of Wisconsin. https://babel.hathitrust.org/cgi/pt?id=mdp.39015019387169&view=1up&seq=9 .

HERNÁNDEZ, Clara y José A. SAMPER, 2003: "Los dialectalismos en el léxico disponible de Gran Canaria. Análisis de un centro de interés", en Francisco MORENO FERNÁNDEZ, Francisco GIMENO, José A. SAMPER, María Luz GUTIÉRREZ, María VAQUERO y César HERNÁNDEZ (coord.), *Lengua, Variación y contexto*, vol. I, Madrid: Arco/Libros, 339-353.

HERNÁNDEZ MUÑOZ, Natividad, 2004: *El léxico disponible de los estudiantes conquenses*, Salamanca: Ediciones Universidad de Salamanca.

HERNÁNDEZ MUÑOZ, Natividad, 2006: *Hacia una teoría cognitiva integrada de la disponibilidad léxica: El léxico disponible de los estudiantes castellano-manchegos*. Tesis Doctoral, Ediciones Universidad de Salamanca.

HERNÁNDEZ MUÑOZ, Natividad, 2009: "Aspectos sociolectales del léxico dialectal", *Spanish in Context* 6:2, 224-248.

HERNÁNDEZ MUÑOZ, Natividad, 2009: "Variación léxica y zonas dialectales de Castilla-La Mancha", *Revista de Filología Española (RFE)* LXXXIX 2, 279-300.

HERNÁNDEZ MUÑOZ, Natividad, y Carmela TOMÉ CORNEJO, 2017: "Léxico disponible en primera y segunda lengua: bases cognitivas", *Palabras Vocabulario Léxico*, 99-122.

HERRANZ, Cristina, 2018: "Disponibilidad léxica de los futuros profesores de Educación Infantil y Primaria", *Revista Electrónica Interuniversitaria de Formación del profesorado* 21(1), 143-159.

HERRANZ, Cristina, 2020: *Palabra de maestro. Análisis del léxico disponible de los futuros docentes*. Frankfurt: Peter Lang.

HERRANZ, Cristina y Miguel MARCOS, 2019: "Análisis del léxico disponible español de extranjeros que estudian los grados de educación", *OGIGIA* 26, 5-30 https://revistas.uva.es/index.php/ogigia/article/view/3702/2983

HERRERA, Honesto, Rosario MARTÍNEZ, y Marian AMENGUAL, 2011: *Estadística aplicada a la investigación lingüística*, Madrid: EOS Universitaria.

HIGALGO GALLARDO, Matías, 2017: "Evolución diacrónica del léxico disponible de estudiantes sinohablantes de ELE", S*inoELE Revista de enseñanza de español a hablantes de chino* 16, 1-16.

HIROTO, Ueda, 2023: Varilex. https://h-ueda.sakura.ne.jp/varilex/index.html

IBÁÑEZ, Romualdo, Moncada, F., Cornejo, F. y Arriaza, V, 2017: "Los géneros del conocimiento en textos escolares de educación primaria", *Calidoscopio* 15, *3*, 462-476.

INSTITUTO DE ESTADÍSTICA DE LA UNESCO, 2013: *Clasificación Internacional Normalizada de la Educación CINE 2011*, Canadá: Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO).

JIMÉNEZ BERRIO, Felipe, 2019: *Estudio sociolingüístico del léxico disponible de escolares navarros*, Pamplona: EUNSA.

JOSSELSON, Harry, 1953: *Russian Words Count*. Detroit: Wayne University.

JUÁREZ, María, 2019: "Influencia de la formación inicial d*el profesor*ado en los hábitos lectores y en el concepto de educación literaria", *Investigaciones Sobre Lectura (ISL)* 12, 99-115.

JUILLAND, A. y E. CHANG-RODRÍGUEZ, 1964: *Frequency Dictionary of Spanish Words*, La haya: Mouton.

JUSTO HERNÁNDEZ, Hortensia, 1987: *Disponibilidad léxica en colores*. Tesis de Maestría, Universidad Nacional Autónoma de México en México D.F.

KÄDING, J. W., 1897: *Häufigkeitswörterbuch der Deutschen Sprache*, Berlin: Steglitz.

KLOSS, Steffanie y Angie QUINTANILLA, 2022: "Disponibilidad léxica para el concepto de escritura en estudiantes de periodismo y periodistas", *Formación Universitaria* 15, 2, 3-10.

KUMAR, Ranjit, 2011: *Research Methodology a step-by-step guide for beginners*, Londres: Sage.

LAGÜÉNS GRACIA, Vicente, 2008: "La variable *sexo* en el léxico disponible de los jóvenes aragonenses", *Estudios sobre disponibilidad léxica en los jóvenes aragoneses*, Actas de las «Jornadas sobre disponibilidad léxica en los jóvenes aragoneses» celebradas en Zaragoza del 9 al 10 de noviembre de 2005, Zaragoza: Institución Fernando el Católico (IFC), 103-162.

LAKOFF, George y Mark JOHNSON, 1980: *Metaphor We Life by*. Chicago: Chicago University Press.

LARA, Luis Fernando, 2006: *Curso de lexicología*, México, D.F.: El Colegio de México.

LABOV, William. 1966. *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistic.

LABOV, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania.

LARRAÑAGA, Elisa, YUBERO Santiago y Pedro CERRILLO, 2008: *Estudio sobre los hábitos lectores de los universitarios españoles*, CEPLI Fundación SM.

LIN, Jin, 2012: "El estudio de disponibilidad léxica de los estudiantes chinos de español como lengua extranjera", *MarcoELE* 14 https://marcoele.com/descargas/14/lin-disponibilidad_lexica.pdf

LÓPEZ CHÁVEZ, Juan, 1992: "Alcances panhispánicos del léxico disponible", *Lingüística* 4, 26-124.

LÓPEZ CHÁVEZ, Juan, 1994: "Comportamiento sintáctico de algunos verbos ordenados según su grado de disponibilidad léxica", *Revista de Estudios de Adquisición de la lengua española REALE* 1, 67-84.

LÓPEZ CHÁVEZ, Juan y STRASSBURGER Carlos, 1987: "Otro cálculo del índice de disponibilidad léxica". *Actas del IV Simposio de la Asociación Mexicana de Lingüística Aplicada. Presente y perspectivas de la lingüística computacional en México*. México: UNAM.

LÓPEZ CHÁVEZ, Juan y STRASSBURGER Carlos, 1991: "Un modelo para el cálculo del índice de disponibilidad léxica individual". En Humberto LÓPEZ MORALES (ed.) *La enseñanza del español como lengua materna. Actas del II Seminario sobre "Aportes de la lingüística a la enseñanza de la lengua materna"*. 91-112. Río Piedras: Universidad de Puerto Rico.

LÓPEZ CHÁVEZ, Juan, y Carlos STRASSBURGER, 2000: "El diseño de una fórmula matemática para obtener un índice de disponibilidad léxica confiable", *Anuario de Letras* 30, 227-251.

LÓPEZ FERRERO, Carmen, y Sergio TORNER CASTELLS, 1999: "Disponibilidad léxica y ponderación en el discurso académico: el uso de los adjetivos en el corpus 92", *Revista de Estudios de Adquisición de la lengua española REALE* 11, 23-45.

LÓPEZ MORALES, Humberto, 1973: *Disponibilidad léxica de los escolares de San Juan*. Ms.

LÓPEZ MORALES, Humberto, 1994a: "Índices de complejidad sintáctica y memoria inmediata", *Revista de Estudios de Adquisición de la lengua española REALE* 1, 85-105.

LÓPEZ MORALES, Humberto, 1994b: *Métodos de investigación lingüística*, Salamanca: Ediciones Colegio de España.

LÓPEZ MORALES, Humberto, 1995-1996: "Los estudios de disponibilidad léxica: pasado y presente", *Boletín de Filología* XXXV, 245-259.

LÓPEZ MORALES, Humberto, 1999: *Léxico disponible de Puerto Rico*, Madrid: Arco Libros.

LÓPEZ MORALES, Humberto y Ester TRIGO IBÁÑEZ, 2019: "La disponibilidad léxica: tendencias actuales y perspectivas de futuro", *OGIGIA-Revista Electrónica de Estudios Hispánicos* 25, 7-10.

LÓPEZ MORALES, Humberto, y Francisco Joaquín GARCÍA MARCOS, 1995: "Disponibilidad léxica en Andalucía proyecto de investigación", *Revista de Estudios de Adquisición de la lengua española REALE* 3, 65-76.

LORGE, Irving, 1952: "Prefacio". En RODRÍGUEZ Bou (Ed.), *Recuento de Vocabulario Español. XV-XVI*. Río Piedras: Universidad de Puerto Rico.

LUGO DE USATEGUI, Kenia, 2005: "El proceso de lectura de hipertextos: ¿Una nueva forma de leer?", *Educere* Vol.9 30, 365-372.

LORÁN, Roberto, 1983: *Un índice de disponibilidad léxica*. Tesis de Maestría, Universidad de Puerto Rico en Mayagüez.

LORÁN, Roberto y Humberto LÓPEZ MORALES, 1983: *Nouveau calcul de l'indice de disponibilité*. MS.

LYONS, John, 1997: *Semántica lingüística*, Barcelona: Paidós.

MCENERY, Tony y Andrew HARDIE. 2012: *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

MACKEY, William, 1971: *Le vocabulaire disponible du français*, 2 vols., Paris-Bruxelles-Montréal: Dibier.

MAFOKOZI, Joseph, 2009: *Introducción a la estadística*, Madrid: Editorial CCS.

MAHECHA MAHECHA, Viviana, y Geral Eduardo MATEUS FERRO, 2017: "El léxico disponible y sus mecanismos de asociación: un análisis con grafos", *Palabras Vocabulario Léxico*, 123-142.

MANJÓN-CABEZA, Antonio, 2010: "Aproximación a la organización semántica del léxico sobre Juegos y Diversiones", *ELUA* 24,199-224.

MANRESA, Mireia, 2009: "El hábito lector a través de la voz adolescente: De la vida al texto", *Lectura y Vida. Revista Latinoamericana de Lectura* 30, 4, 32-42.

MANGADO CRUZ, Marta, y María ARETA LARA, 2008: "Procesamiento informático de datos para la elaboración de diccionarios de disponibilidad léxica", *Actas del XXXVII Simposio Internacional de la Sociedad Española de Lingüística (SEL)*, 479-493.

MARCOS-CALVO, Miguel, Cristina HERRANZ y Ricardo-María JÍMENEZ-YÁÑEZ, 2022: "El conocimiento de otras lenguas en el léxico disponible español en estudiantes de secundaria y bachillerato", *TEJUELO. Didáctica de la Lengua y la Literatura. Educación* 35, 3, 43-72 <https://doi.org/10.17398/1988-8430.35.3.43>

MARTÍNEZ, Esther, 2007: *Disponibilidad léxica en las comunidades de habla alicantinas*. Tesis doctoral. Alicante: Universidad de Alicante

MARTÍNEZ-LARA, José Alejandro, 2016: "Corpus de Interacciones de Jóvenes Universitarios: una experiencia para la investigación del lenguaje en contexto", Revista Latinoamericana de Estudios del Discurso 16, 1, 67-81.

MARTÍNEZ-LARA, José Alejandro, 2021: "Incidencia de los años de escolaridad y cantidad de lectura en el léxico disponible de un grupo de estudiantes universitarios del área de pedagogía", *Boletín de filología* 56, 2, 519-548.

MARTÍNEZ-LARA, José Alejandro, 2023: "Léxico venezolano: análisis y propuesta metodológica desde el enfoque de la disponibilidad léxica", en Enrique PATO (ed.), *Estudios sobre el español de Venezuela*. Iberoamericana/Vervuert.

MARTÍNEZ SEGURA, Gricelda, 2012: *Disponibilidad léxica de las/os estudiantes de II de Bachillerato Académico, sección "A" del instituto 1° de mayo de 1954. Una propuesta para la enseñanza del vocabulario*. Tesis de Maestría, Universidad Pedagógica Nacional de Honduras.

MATEO GARCÍA, M.ª Victoria, 1997: *Disponibilidad léxica en el COU almeriense. Estudio de la estratificación social*. Tesis Doctoral, Universidad de Granada.

MATEO GARCÍA, M.ª Victoria, 1998: *Disponibilidad léxica en el COU almeriense. Estudio de la estratificación social*, Almería: Universidad de Almería.

MATEUS-FERRO, Geral Eduardo, Laura Marcela CASTIBLANCO y Pedro Augusto ÁLVAREZ-BERMÚDEZ, 2018: "Mecanismos lógicos y analógicos en la producción del léxico disponible", *FOLIOS* 47, 133-152.

MATHY, Maurice, 1952: *Vocabulaire de base du latín*. París: OCDL.

MENA OSORIO, Mónica. 1986. *Disponibilidad léxica infantil en tres niveles de enseñanza básica*. Tesis de maestría, Concepción: Universidad de Concepción.

MENDÍVIL GIRO, Jose. 2009: "Palabras con estructura externa". En MIGUEL, Elena de (ed.), *Panorama de la lexicología*, 83-112. Barcelona: Editorial Ariel.

MENDOZA PUERTAS, Jorge, 2018: "El léxico disponible de 82 estudiantes coreanos de español como lengua extranjera", *MarcoEle* 26 https://marcoele.com/descargas/26/mendoza-lexico_disponible-coreanos.pdf

MESA CANALES, Rosa, 1989: *Disponibilidad léxica en preescolares*. México, D.F.: Universidad Autónoma de México.

MICHÉA, René, 1950: "Vocabulaire et culture", *Les Langues Modernes* 44, 188-189.

MICHÉA, René,1953: "Mots fréquents et mots disponibles. Un aspect nouveau de la statistique du langage", *Les Langues Modernes* 47, 338-344.

MIGUEL, Elena de (Ed.), 2009: *Panorama de la lexicología*, Barcelona: Editorial Ariel.

MORALES, Amparo, 1986: *Léxico básico del español de Puerto Rico*. San Juan: Academia Puertorriqueña de la Lengua Española.

MORENO FERNÁNDEZ, Francisco, 1990: *Metodología sociolingüística*, Madrid: Gredos.

MORENO FERNÁNDEZ, Francisco, 2012: "Disponibilidad léxica: cuestiones metodológicas: A propósito de disponibilidad léxica de los estudiantes hispanos de Redwood City, CA", *Revista Nebrija de Lingüística Aplicada* 11.

MORENO FERNÁNDEZ, Francisco, José E. MORENO FERNÁNDEZ y Antonio GARCÍA DE LAS HERAS, 1995: "Cálculo de disponibilidad léxica. El programa Lexidisp", *Lingüística 7*, 243-249.

MORENO FERNÁNDEZ, Francisco, 2012: "Disponibilidad léxica: cuestiones metodológicas: A propósito de disponibilidad léxica de los estudiantes hispanos de Redwood City", *Revista Nebrija de Lingüística Aplicada* 11, (SP).

MÜLLER, Charles, 1973: *Estadística lingüística*. Madrid: Gredo.

MUNITA, Felipe, 2014: *El mediador escolar de lectura literaria. Un estudio del espacio de encuentro entre prácticas didácticas, sistemas de creencias y trayectorias personales de lectura*. Tesis doctoral, Universitat Autònoma de Barcelona en Barcelona.

MUNITA, Felipe, 2018: "El sujeto lector didáctico: 'lectores que enseñan y profesores que leen'", *Álabe* 17, 1-19.

MUÑOZ NÚÑEZ, 2010: El uso figurado en algunos casos de unidades plurilexemáticas: colocaciones y compuestos sintagmáticos. *ELUA* 24, 253-270.

MURILLO, Marielos, 1993: "Disponibilidad léxica en los preescolares: estudio de cinco campos semánticos", *Káñina*, 17(2), 117-127.

MURILLO, Marielos, 1994: "Comidas y bebidas: estudio de la disponibilidad léxica en preescolares", *Káñina*, 18(2), 117-133.

MURILLO, Marielos, 1998: "Crecimiento de la disponibilidad léxica: niños de preescolar y primer ciclo de la educación básica costarricense", *Revista de Filología y Lingüística de la Universidad de Costa Rica* XXV, 2, 187-203.

NALESSO, Giulia, 2022: "Disponibilidad léxica y ortografía en ELE: un estudio para la enseñanza de la lengua", *RILEX. Revista sobre investigaciones léxicas*, 5/I, 7-36.

NJOCK, Pierre-Emmanuel, 1978: *L'univers familier de l'enfant africain*. Québec: Centre International de Recherche sur le Bilinguisme. Disponible en https://eric.ed.gov/?id=ED192580

NOVOA LAGOS, Abraham Benjamín, 2013: *Incidencia de la competencia léxica en la comprensión de lectura*. Tesis de maestría, Universidad de Concepción de Chile.

NÚÑEZ, Fredy, 2021: *Diseño y desarrollo de un modelo de desambiguación léxica automática para el procesamiento del lenguaje natural*, Tesis Doctoral, Pontificia Universidad Católica de Chile.

NÚÑEZ DELGADO, M.ª Pilar, y María SANTAMARINA SANCHO, 2014: "Prerrequisitos para el proceso de aprendizaje de la lectura y la escritura: conciencia fonológica y destrezas orales de la lengua", *Lengua y Habla* 18, 72-92.

OSSES, Sonia y Sandra JARAMILLO, 2008: "Metacognición: un camino para aprender a aprender", *Estudios pedagógicos* 34, 1, 187-197. https://dx.doi.org/10.4067/S0718-07052008000100011

OTAOLA, Concepción, 2004: *Lexicología y semántica léxica*, Madrid: Ediciones Académicas.

PACHECO CARPIO, Carmen, Juan Silvio CABRERA ALBERT e Iselys GONZÁLEZ LÓPEZ, 2017: "Incidencia de la variable "sexo" en la disponibilidad léxica de estudiantes de preuniversitario en Pinar del Río, Cuba", *Revista de Lenguaje y Cultura* Vol. 22, 237-253.

PACHECO MIRABAL, Antonio, Selina PONCE-CASTAÑEDA y Salvador PALOMARES-SÁNCHEZ, 2016: "Disponibilidad léxica matemática en estudiantes de ingeniería y ciencias", *Revista Iberoamericana de Educación Matemática* 47, 44-61.

PAREDES GARCÍA, Florentino, 1999: "La ortografía en las encuestas de disponibilidad léxica", *Revista Estudio Adquisición de la Lengua Española (REALE) 11*, 75-98.

PAREDES GARCÍA, Florentino, 2000: "Disponibilidad de los extranjerismos en estudiantes de educación secundaria", *ASELE*, Actas XI, 567-576.

PAREDES GARCÍA, Florentino, 2012: "Desarrollos teóricos y metodológicos recientes de los estudios de disponibilidad léxica", *Revista Nebrija de Lingüística Aplicada* 11, SP.

PAREDES GARCÍA, Florentino, 2014: "A vueltas con la selección de 'centros de interés' en los estudios de disponibilidad léxica: para una propuesta renovadora a propósito de la disponibilidad léxica en ELE", *Revista Nebrija de Lingüística Aplicada*, 16.

PAREDES GARCÍA, Florentino, y Diego GALLEGO GALLEGO, 2019: "Procedimiento neológicos en el léxico disponible de español como lengua materna y como lengua extranjera", *Ogigia. Revista electrónica de estudios hispánicos* 25, 109-138.

PARODI, Giovanni. 2004: "Textos de especialidad y comunidades discursivas técnico-profesionales: una aproximación basada en corpus computarizado". *Estudios filológicos*, 39, 7-36. https://dx.doi.org/10.4067/S0071-17132004003900001

PARODI, Giovanni. 2008: "Lingüística de corpus: una introducción al ámbito", *Revista de Lingüística Teórica y Aplicada* 46, 1. 93-119.

PARRADO COLLANTES, Milagrosa, Manuel Francisco ROMERO OLIVA y Ester TRIGO IBÁÑEZ, 2018: "La experiencia literaria en la formación de futuros docentes: el viaje iniciático de nuestras biografías lectoras en 10 hashtag" en AMAR RODRÍGUEZ, Víctor Manuel (Ed.): *Miradas y voces de futuros maestros*, Barcelona: Octaedro, 57-84.

PAVEY, Emma: 2010. *The Structure of Language. An Introduction to Grammatical Analysis*. Cambridge: Cambridge University Press.

PENG, S. y C. MORENO. 2021: "Estrategias de cortesía y valoraciones negativas en el comercio electrónico. Un estudio contrastivo chino-español". *Revista Española de Lingüística Aplicada*, 34(2). p. 585 – 610. https://doi.org/10.1075/resla.19024.

PEREA, Manuel, y Eva ROSA, 1999: "Psicología de la lectura y procesamiento léxico visual: Una revisión de técnicas experimentales y de procedimientos de análisis", *Psicológica* 20, 65-90.

PÉREZ, Francisco, 2005: *Pensar y hacer el diccionario*. Caracas: Los libros de El Nacional.

PIERA, Carlos, 2009: "Una idea de la palabra", en Elena de MIGUEL (ed.), *Panorama de la lexicología*. Ariel, Barcelona, 25-49.

PORTO DAPENA, José-Álvaro. 2002: *Manual de técnica lexicográfica*. Madrid: Arco/Libros.

PRADO ARAGONÉS, Josefina, y M.ª Victoria GALLOSO CAMACHO,2008: "La variable sexo en el léxico disponible de alumnos de primaria y bachillerato de Huelva: resultados cuantitativos", en Blas, J. L. *et al*. (ed), *Discurso y sociedad II. Nuevas contribuciones al estudio de la lengua en contexto*. 583-595. Universidad Jaume I.

PRADO ARAGONÉS, Josefina, y M.ª Victoria GALLOSO CAMACHO, 2015: *El léxico disponible de Extremadura y comparación con el de Andalucía*, Huelva: Universidad de Huelva.

PRADO ARAGONÉS, Josefina, M.ª Victoria GALLOSO CAMACHO y Manuel CÉLIO CONCEIÇÃO, 2009: *La disponibilidad léxica en situación de contacto de lenguas en las zonas limítrofes de Andalucía y Extremadura (España) y Algarve y Alentejo (Portugal)*, Huelva: Universidad de Huelva.

PRESEEA, 2023: https://preseea.uah.es/

QUEZADA, Camilo, 2007: "Potencia estadística, sensibilidad y tamaño de efecto: ¿un nuevo canon para la investigación?" *Onomázein* 2, 16, 159-17.

QUINTANILLA ESPINOZA, Angie, y Pedro SALCEDO LAGOS, 2019: "Disponibilidad léxica en procesos de formación inicial de futuros profesores de inglés", *Rev. Bras. Lingüíst. Apl.* Vol. 19, 529-554.

RAMÍREZ, Yessenia, 2019: "Una aproximación a la complejidad del caudal léxico de profesores chilenos en formación: análisis cualitativo y cuantitativo de estructuras plurilexicales". *Nueva Revista del Pacífico* 71, 143-160.

RAMÍREZ LEYVA, Elsa M., 2009: "¿Qué es leer? ¿Qué es la lectura?", *Investigación Bibliotecológica Vol.*23, n°47, 161-188.

REAL ACADEMIA ESPAÑOLA y ASOCIACIÓN DE ACADEMIAS DE LA LENGUA, 2005: *Diccionario Panhispánico de Dudas*. Madrid: Espasa.

REAL ACADEMIA ESPAÑOLA, 2009: *Nueva Gramática de la Lengua Española*. Madrid, Espasa-Calpe.

REAL ACADEMIA ESPAÑOLA y ASOCIACIÓN DE ACADEMIAS DE LA LENGUA, 2014: *Diccionario de la Lengua Española*. Madrid: Espasa.

RIFFO OCARES, Bernardo, Fernando REYES REYES, Abraham NOVOA LAGOS, Mónica VÉLIZ DE VOS y Ginette CASTRO YÁÑEZ, 2014: "Competencia léxica, comprensión lectora y rendimiento académico en estudiantes de enseñanza media", *Literatura y Lingüística* 30, 165-180.

RÍOS GONZÁLEZ, Gabriel, 2007: "Diferencias léxicas entre el hombre y la mujer en tres centros de interés: saludos, temas de conversación y despedidas", *Filología y Lingüística* XXXIII 1, 151-166.

RODRÍGUEZ MUÑOZ, Francisco José, e Isabel Ofelia MUÑOZ HERNÁNDEZ, 2009: "De la disponibilidad a la didáctica léxica", *Tejuelo* 4, 8-18.

RODRÍGUEZ ROMERO, María. 1998: "El cambio educativo y las comunidades discursivas: representando el cambio en tiempos de postmodernidad", *Revista de Educación*, 317, 157-184.

ROJAS, Carlos, 2020: *Perfil léxico de adultos mayores chilenos de tercera y cuarta edad. Estudio transeccional*. Tesis doctoral. Concepción, Universidad de Concepción.

ROJAS, Darío, Carolina del Carmen ZAMBRANO y Pedro SALCEDO, 2017: "Metodología de análisis de disponibilidad léxica en alumnos de pedagogía a través de la comparación jerárquica de lexicones", *Formación Universitaria* Vol. 10, 3-14.

ROJAS DÍAZ, Darío, Carolina ZAMBRANO MATAMALA, Pedro SALCEDO LAGOS y Miguel FRIZ CARRILLO, 2018: "Un enfoque de reconocimiento de patrones para el análisis de disponibilidad léxica en estudiantes de pedagogía en matemáticas", *Estudios Filológicos* 62, 333-358.

ROJO, Guillermo. 2021: *Introducción a la lingüística de corpus en español*. Oxon: Routledge.

ROMÁN, Belén, 1985: *Disponibilidad léxica de escolares de Dorado, Puerto Rico*. Tesina inédita, Universidad de Puerto Rico en Río Piedra.

ROMERO OLIVA, Manuel Francisco, y Ester TRIGO IBÁÑEZ, 2019: "Entre la realidad y la experiencia en la formación de nuevos lectores. Un análisis del discurso de especialistas más allá de la propia teoría" en TATOJ, Cecylia (Ed.): *Voces y caminos en la enseñanza de español/LE desarrollo de las identidades en el aula*, 119-137.

ROMERO-PÉREZ, Ivón, ALARCÓN-VÁSQUEZ, Yolima y GARCÍA-JIMÉNEZ, Rafael. (2018). Lexicometría: enfoque aplicado a la redefinición de conceptos e identificación de unidades temáticas. *Biblios*, (71), 68-80. https://dx.doi.org/10.5195/biblios.2018.466

ROSCH, Eleanor, 1978: "Principles of categorization", en ROSCH, Eleanor y Barbara LLOYD (eds.): *Cognition and categorization*, Hillsdale (N.J.): Erlbaum, 27-48.

ROSSELLI, Mónica, Esmeralda MATUTE y Alfredo ARDILA, 2006: "Predictores neuropsicológicos de la lectura en español", *Revista de Neurología* 42 (4), 202-210.

RUBIO SÁNCHEZ, Roberto, 2017: "Acercamiento al léxico disponible de 173 estudiantes italianos preuniversitarios de español como lengua extranjera", *Palabras Vocabulario Léxico*, 143-161.

RUÍZ BASTO, Araceli, 1987: *Disponibilidad léxica de los alumnos de primer ingreso en el colegio de ciencias y humanidades. Plantel Naucalpan*. Tesis de Licenciatura en Lengua y Literatura Hispánica. Universidad Autónoma de México.

SAINE CAMARGO, Ana María, 2008: *El léxico disponible de los estudiantes de la escuela media bonaerense: aspectos metodológicos y sociolingüísticos*. Tesis de doctorado. Universidad Nacional de Educación a Distancia.

SALCEDO LAGOS, Pedro, María DEL VALLE, Anita FERREIRA y Roberto HERNÁNDEZ, 2008: "Léxico disponible de estudiantes de Concepción Chile en las áreas de datos y azar, álgebra, número y geometría" [https://bit.ly/3UocOAU, fecha de consulta: 08 de enero de 2023].

SALCEDO LAGOS, Pedro, Oscar NAIL KROYER y Carla ARZOLA ZAPATA, 2012: "Análisis de relaciones semánticas del léxico disponible en matemáticas en un hipermedio adaptativo", *Nuevas Ideas en Informática Educativa, TISE*, 160-164.

SALCEDO LAGOS, Pedro, y María DEL VALLE LEO, 2013: "Disponibilidad léxica matemática en estudiantes de enseñanza media de Concepción, Chile", *Atenas. Revista Científico-Pedagógica* Vol.4, 1-16.

SAMPER PADILLA, José Antonio, 1998a: *Léxico del habla culta de Las Palmas de Gran Canaria*, Las Palmas de Gran Canaria: Cabildo Insular de Gran Canaria y Universidad de Las Palmas de Gran Canaria.

SAMPER PADILLA, José Antonio, 1998b: "Criterios de edición del léxico disponible: sugerencias", *Lingüística* Vol.10, 311-332.

SAMPER PADILLA, José Antonio, 2021: "Disponibilidad léxica y sociolingüística", en Maribel Serrano y M[a.] Ángeles Calero (Eds.), *Aplicaciones de la disponibilidad léxica*. Sevilla: tirant humanidades, 173-206.

SAMPER PADILLA, José Antonio y HERNÁNDEZ Clara, 1997: "El estudio de la disponibilidad léxica en gran canaria: datos iniciales y variación sociolingüística", En J. DORTA y M. ALMEIDA (ed.) *Contribuciones al estudio de la lingüística hispánica. Homenaje al Prof. Ramón Trujillo*, Barcelona: Montesinos, 229- 239.

SAMPER PADILLA, José Antonio y HERNÁNDEZ Clara, 2006: "Densidad de dialectalismos y condicionantes sociales en el léxico disponible de Canarias", en Mercedes SEDANO, Adriana BOLÍVAR y Martha SHIRO (coord.), *Haciendo lingüística. Homenaje a Paola Bentivoglio*, Caracas: Comisión de estudios de Postgrado Facultad de Humanidades y Educación: Universidad Central de Venezuela.

SÁNCHEZ, Reinaldo, 2015: "t-Student. Usos y abusos", *Revista mexicana de cardiología* 26, 1, 50-61.

SÁNCHEZ, Víctor y Murillo Marielos, 2006: *Disponibilidad léxica de los niños preescolares costarricenses*, San José: Universidad de Costa Rica.

SÁNCHEZ-SAUS, Marta, 2011: *Bases semánticas para el estudio de los centros de interés léxico disponible. Disponibilidad léxica de informantes extranjeros en las universidades andaluzas*. Tesis Doctoral, Universidad de Cádiz.

SÁNCHEZ-SAUS, Marta, 2012: "Fundamentos historiográficos de los centros de interés del léxico disponible. Los tipos de contenido léxico de la semántica histórica", *Historiografía lingüística: líneas actuales de investigación*, 800-808.

SÁNCHEZ-SAUS, Marta, 2013: "Apuntes para una caracterización semántica de los centros de interés en los estudios de disponibilidad léxica", *Anejo LXXXVI de Analecta Malacitana*, 235-252.

SÁNCHEZ-SAUS, Marta, 2016: *Léxico disponible de los estudiantes de español como lengua extranjera en las universidades andaluzas*, Sevilla: Editorial Universidad de Sevilla.

SÁNCHEZ-SAUS, Marta, 2019: *Centros de interés y capacidad asociativa de las palabras*, Sevilla: Editorial Universidad de Sevilla.

SÁNCHEZ-SAUS, Marta, 2022: "Redes semánticas, léxico disponible y didáctica del vocabulario en ELE: un análisis por niveles de español", *Tejuelo* 35, 3, 167-204.

SANDU, Bianca, 2012: "La disponibilidad léxica en alumnos rumanos de ELE: incidencia de la variable 'sexo/género' y su correlación con el 'nivel escolar'", *Lingua Americana* Año XVI 31, 61-85.

SANDU, Bianca, 2013: "La disponibilidad léxica en alumnos rumanos de español como lengua extranjera", *Estudios interlingüísticos* 1, 121-133.

SANKOFF, David. 1978. *Linguistic variation: Models and methods*. New York: Academic Press.

SANTOS DÍAZ, Inmaculada Clotilde, 2015a: "Análisis comparativo del léxico en español y en lengua extranjera del futuro profesorado", *Revista Digital E-Aesla* 1, https://cvc.cervantes.es/lengua/eaesla/pdf/01/67.pdf

SANTOS DÍAZ, Inmaculada Clotilde, 2015b: *Evaluación de la competencia léxica bilingüe en estudiantes del Máster Universitario en Profesorado. Análisis de pruebas de disponibilidad léxica y de identificación de tecnicismos en español, inglés y francés*. Tesis Doctoral, Universidad de Málaga.

SANTOS DÍAZ, Inmaculada Clotilde, 2016: "La influencia del centro educativo en el léxico disponible en lengua extranjera", en AMOR ALMEDINA, María Isabel, Juan Luis LUENGO ALMENA y María MARTÍNEZ ATIENZA (eds.): *Educación intercultural metodología de aprendizaje en contextos bilingües*, España: Atrio.

SANTOS DÍAZ, Inmaculada Clotilde, 2017a: "Incidencia de la lectura en el vocabulario en lengua materna y extranjera", *Ocnos. Revista de Estudios sobre lectura* 16(1), 79-88.

SANTOS DÍAZ, Inmaculada Clotilde, 2017b: "Selección del léxico disponible: propuesta metodológica con fines didácticos", *Portal Linguarum* 27, 122-139.

SANTOS DÍAZ, Inmaculada Clotilde, 2017c: "Organización de las palabras en la mente en lengua materna y lengua extranjera (inglés y francés)", *Pragmalingüística* 25, 603-617.

SANTOS DÍAZ, Inmaculada Clotilde, 2020: *El léxico bilingüe del futuro profesorado*, Berlín: Peter Lang.

SANTOS DÍAZ, Inmaculada Clotilde, María JUÁREZ CALVILLO y Ester TRIGO IBÁÑEZ, 2021: "Motivación por la lectura académica de futuros docentes", *Educação & Formação* Vol. 6, n°1, 1-21.

SANTOS DÍAZ, Inmaculada Clotilde, y María JUÁREZ CALVILLO, 2022: "El concepto de educación del futuro profesorado desde la disponibilidad léxica según su formación académica", *Tejuelo* 35 (3), 263-298.

SANTOS PALMOU, Xandra, 2017: "El vocabulario fundamental: historia, definición y nuevas propuestas aplicadas a la enseñanza de ELE". *E-Aesla* 3. 110-120.

SARALEGUI, Carmen, y Cristina TABERNERO, 2008: "Aportación al proyecto panhispánico de léxico disponible: Navarra", *Actas del XXXVII Simposio Internacional de la Sociedad Española de Lingüística (SEL)*, 745-761.

SCHIEFELE, Ulrich, Ellen SCHAFFNER, Jens MÖLLER y Allan WIGFIELD, 2012: "Dimensions of Reading motivation and their relation to Reading behavior and competence", *Reading Research Quarterly* 47(4), 427-463.

SEBASTIÁN, Nuria, Mª. Antonia MARTI, Manuel CARREIRAS y Fernando CUETOS, 2000: *LEXESP: Léxico informatizado del español*. Universidad de Barcelona, Barcelona.

SEDANO, Mercedes. 2011. Manual de gramática del español, con especial referencia al español de Venezuela. Caracas: UCV-CEP-CDCH.

SEOANE, T, J. MARTÍN, E. MARTÍN-SÁNCHEZ, S. LURUEÑA-SEGOVIA y F. ALONSO MORENO, 2007: "Estadística: Estadística Descriptiva y Estadística Inferencial", *SEMERGEN* 33, 9, 466-71

SERFATI, Mohamed, 2017: "Incidencia cuantitativa del factor 'lengua materna' en la disponibilidad léxica de estudiantes marroquíes de español como Lengua Extranjera (nivel universitario)", *Estudios interlingüísticos* 5, 121-145.

SERRANO ZAPATA, Maribel, y & M.ª Ángeles CALERO FERNÁNDEZ (Eds.), 2021: *Aplicaciones de la disponibilidad léxica*, Valencia: Tirant humanidades.

SERRANO ZAPATA, Maribel, 2004: "Aspectos sociolingüísticos del léxico disponible castellano de los preuniversitarios leridanos", *Pragmalingüística* 12, 147-165.

SILVA-CORVALÁN, Carmen y Andrés HENRÍQUEZ-ARIAS, 2017: *Sociolingüística y pragmática del español*. Washington, DC: Georgetown University Press.

THARP, J., 1939: The Basic French Vocabulary. New York: Holt.

THORNDIKE, E., 1921: *Teacher's Word Book*, New York, MacMillan.

TOMÉ CORNEJO, Carmela, 2015: *Léxico disponible. Procesamiento y aplicación a la enseñanza de ELE*. Tesis Doctoral, Universidad de Salamanca.

TORRES PERDOMO, María Electa, 2003: "La lectura. Factores y actividades que enriquecen el proceso", *Educere* Vol.6, n°20, 389-396.

TRIGO IBÁÑEZ, Ester, 2007: *El léxico disponible de la provincia de Sevilla: variación versus Déficit*. Tesis doctoral inédita, Universidad de Cádiz en Cádiz.

TRIGO IBÁÑEZ, Ester, 2011: *Dialectología y cultura. El léxico disponible de los preuniversitarios sevillanos*, Valencia: Aduana Vieja.

TRIGO IBÁÑEZ, Ester, y GONZÁLEZ MARTÍNEZ, Adolfo, 2011: "Estudio del comportamiento de la variable sexo en el léxico disponible de los preuniversitarios sevillanos", *Diálogos de la Lengua* III, 28-41.

TRIGO IBÁÑEZ, Ester e Inmaculada Clotilde SANTOS DÍAZ, 2021: "Influencia del tipo de centro educativo (público y privado) sobre el léxico disponible de los preuniversitarios sevillanos", en Maribel Serrano y Mª. Ángeles Calero (Eds.), *Aplicaciones de la disponibilidad léxica*. Sevilla: tirant humanidades, 231-250.

TRIGO IBÁÑEZ, Ester, Manuel Francisco ROMERO OLIVA, e Inmaculada Clotilde SANTOS DÍAZ, 2018: "Elaboración de un corpus cacográfico desde la disponibilidad léxica en estudiantes sevillanos. Un análisis para la enseñanza de la lengua", *Revista de Lingüística y Lenguas Aplicadas* Vol.13, 119-131.

TRIGO IBÁÑEZ, Ester, Manuel Francisco ROMERO OLIVA, e Inmaculada Clotilde SANTOS DÍAZ, 2019: "Aproximación al léxico gastronómico dialectal andaluz desde los repertorios de disponibilidad léxica para una propuesta didáctica", *Verba Hispánica* XXVII, 115-130. Doi: 10.4312/vh.27.1.115-130

TRIGO IBÁÑEZ, Ester, Manuel Francisco ROMERO OLIVA e Inmaculada Clotilde SANTOS DÍAZ, 2020: "Disponibilidad léxica y dominio de la ortografía: un estudio empírico basado en la influencia de los factores sociales", *Onomázein* 47, 27-45.

TRIGOS CARRILLO, Lina Marcela, 2012: *¿Ensayamos? Manual de redacción de ensayos*, Bogotá: Editorial Universidad del Rosario.

ULLMAN, Stephen, 1976: *Semántica*. Madrid: Aguilar.

UNESCO-UIS, 2013: *Clasificación Internacional Normalizada de la Educación*. Unesco-UIS.

URZÚA, Paula, Katia SÁEZ, y Max ECHEVERRÍA, 2006: "Disponibilidad léxica matemática. Análisis cuantitativo y cualitativo", *Revista de Lingüística Teórica y Aplicada RLA* 44, 59-76.

URZÚA CARMONA, Paula Carolina, 2018: *Disponibilidad léxica en adultas mayores de la ciudad de concepción: estudio descriptivo*. Tesis Doctoral, Universidad de Concepción.

UZCÁTEGUI, Elisa, 1992: "La sociolingüística de Basil Bernstein y sus implicaciones en el ámbito escolar". *Revista de Educación*, 298: 163-197.

VALENCIA, Alba, 1994: *El léxico de los estudiantes de 4º año de Educación Media. Centro de Interés Procesos mentales*. Santiago de Chile: Serie documentos de estudio 26. CPEIP.

VALENCIA, Alba, 1997: "Disponibilidad léxica. Muestreo y estadísticos", *Onomazein* 2, 197-226.

VALENCIA, Alba, 1998-1999: "Léxico estudiantil. Comentario sobre un repertorio", *BFUCh* XXXVII, 1211-1221.

VALENCIA, Alba, 2010: "Léxico del Color en Santiago de Chile", *Revista de Lingüística Teórica y Aplicada* 48, 2, 141-161.

VALENCIA, Alba, 2011: "Disponibilidad léxica en Aragón y Chile. Revisión contrastiva", *Archivo de Filología Aragonesa (AFA)* 67, 173-200.

VALENCIA, Alba, y Max ECHEVERRÍA, 1999: *Disponibilidad léxica en estudiantes chilenos*, Santiago: Ediciones Universidad de Chile – Universidad de Concepción.

VALENZUELA, Javier, Iraide IBARRETXE-ANTUÑANO y Joseph HILFERTY, 2012: "La semántica cognitiva", en Iraide IBARRETXE-ANTUÑANO y Javier VALENZUELA (eds.) *Lingüística Cognitiva*, Barcelona, Anthropos, 41-68.

VALENZUELA, Marco, Victoria PÉREZ VILLALOBOS, Claudio BUSTOS, y Pedro SALCEDO LAGOS, 2018: "Cambios en el concepto aprendizaje de estudiantes de pedagogía: análisis de disponibilidad léxica y grafos", *Estudios Filológicos* 61, 143-173.

VALLS, Rosa, Marta SOLER, y Ramón FLECHA, 2008: "Lectura dialógica: interacciones que mejoran y aceleran la lectura", *Revista Iberoamericana de Educación* 46, 71-87.

VANDER BEKE, George, 1935: "French word book". *Publications of the American and Canadian Committees on Modern Languages* 50, Canada: University of Toronto. https://babel.hathitrust.org/cgi/pt?id=mdp.39015034357353&view=1up&seq=27

VARLEÉ, León, 1954: *Basis-Woordenboek voor de Franse Taal*. Amsterdam: Neulenhoff.

VÁZQUEZ-CANO, Esteban, Santiago MENGUAL-ANDRÉS y Rosabel ROIG-VILLA, 2015: "Análisis lexicométrico de la especificidad de la escritura digital del adolescente en Whatsapp". *Revista de Lingüística Teórica y Aplicada* 53 (1), 83-105.

VENEGAS, René, ZAMORA, S. y GALDAMES, A.: 2016. "Hacia un modelo retórico-discursivo del macrogénero Trabajo Final de Grado en Licenciatura", *Revista Signos. Estudios de Lingüística* 49, S1, 247-279.

VICTERY, John, 1971: *A study of lexical availability among monolingual-bilingual speakers of spanish and english*. Thesis Master of Arts. Houston, Rice University.

WEST, M. y J. ENDICOTT, 1941: *The New Method English Dictionary Explaining the Meaning of 24.000 items wthin a Vocabulary of 1490 Words*. Paris: Didier.

WINGEYER, Hugo Roberto, 2007: *Léxico disponible de resistencia*. Tesis de doctorado. Universidad de Alcalá.

WINGEYER, Hugo Roberto, 2014: "Disponibilidad léxica de la región Nea de Argentina. Análisis socio-semiótico de formas asociadas al acto de insultar", *Contextos: Estudios de humanidades y ciencias sociales*, 31, 129-141.

YUS, F. 2001: *Ciberpragmática. El uso del lenguaje en Internet*. Ariel.

YUS, F. 2010: *Ciberpragmática 2.0. Nuevos usos del lenguaje en internet*. Barcelona: Ariel.

ZAMBRANO MATAMALA, Carolina, 2021: "Un estudio de la disponibilidad léxica en el ámbito de autorregulación del aprendizaje en la formación inicial docente", *Lingüística y Literatura* 79, 11-33.

ZHOU, Guxin, 2021: *Léxico disponible de estudiantes chinos universitarios de español como lengua extranjera*. Tesis doctoral. Universidad de Alcalá.

ZIPF, George, 1946: *Human Behavior and the Principle of Least Effort*. Cambridge: Mass, Addison-Wesley.

Anexo 1. Encuesta sociológica

ENCUESTA SOCIOLÓGICA

La información suministrada en este cuestionario es totalmente confidencial y anónima, y será utilizada solo con fines académicos. En este contexto, te solicitamos que sea lo más honesto/a posible con tus respuestas. De antemano, agradecemos tu valiosa colaboración.

DATOS PERSONALES

1) Sexo: Hombre ☐ Mujer ☐

2) Edad: _____

3) Lugar de nacimiento (ciudad, país): _____

4) Ciudad actual de residencia: _____

5) ¿Cuál es tu primera lengua? _____

6) ¿Hablas otra lengua de manera fluida? Sí ☐ No ☐

7) ¿Cuál lengua? _____

DATOS DE LOS PADRES

8) País de nacimiento del padre: _____

9) Lengua materna del padre: _____

10) Nivel educativo del padre:

Sin educación formal	☐	Educación universitaria incompleta	☐
Educación Básica	☐	Educación universitaria completa	☐
Educación Media	☐	Posgrado	☐
Educación técnica completa	☐	No sé	☐

11) En caso tal, ¿cuál es la profesión del padre? _____

12) ¿Cuál es su ocupación actual? _____

13) País de nacimiento de la madre: _____

14) Lengua materna de la madre: _____

15) Nivel educativo de la madre:

Sin educación formal	☐	Educación universitaria incompleta	☐
Educación Básica	☐	Educación universitaria completa	☐
Educación Media	☐	Posgrado	☐
Educación técnica completa	☐	No sé	☐

16) En caso tal, ¿cuál es la profesión de la madre? _____

17) ¿Cuál es su ocupación actual? _____

18) En promedio, ¿cuál es la renta mensual de tu familia?

Menos de 314 mil	☐	1.663.000 a 2.899.000	☐
De 314 mil a 546 mil	☐	2.900.000 a 5.057.000	☐
547 mil a 953 mil	☐	Más de 5.057.000	☐
954 mil a 1.662.000	☐	No sé	☐

INFORMACIÓN EDUCATIVA DEL ENCUESTADO/A

19) ¿Dónde realizaste los estudios de educación media?

Colegio municipal ☐ Colegio particular subvencionado ☐

Colegio particular ☐ Otro (especifique): _____

20) ¿En qué año egresaste de la educación media? _____

21) ¿En qué año ingresaste a la universidad (carrera actual)? _____

22) Universidad en la que cursa carrera actualmente:

_____

23) Carrera que cursa:

Pedagogía en Educación Básica	☐	Pedagogía en Educación Parvularia	☐
Pedagogía en Educación Media	☐	Pedagogía en Educación Diferencial	☐
Pedagogía en Inglés	☐	Pedagogía en Educación Física	☐
Pedagogía en Matemática	☐	Letras	☐

Otra mención u otra carrera (especifique): _____

24) Año que cursa en la carrera actual: Primer año ☐ Cuarto año ☐

Otro año (especifique): _____

25) Financiamiento de la carrera universitaria:

Padres/familia ☐

Gratuidad ☐

Beca ¿cuál y porcentaje? _____

Otro (especifique) _____

26) ¿Has cursado estudios técnicos o universitarios previamente a tu carrera actual?

Sí ☐ No ☐ Actualmente curso otra carrera en paralelo ☐

27) En caso tal, ¿qué carrera o estudio previo realizaste o cursas actualmente en paralelo?

_____

28) Si fuera el caso, ¿cuál es tu nivel de estudio técnico o universitario previo?

Incompleto ☐ Completo ☐ En curso ☐ No aplica ☐

Muchas gracias

Anexo 2. Encuesta sobre prácticas lectoras

ENCUESTA SOBRE LA LECTURA

La información suministrada en esta encuesta es totalmente confidencial y anónima. Los datos serán utilizados solo con fines académicos. Tu participación es voluntaria y no repercutirá en tu desempeño académico. Esta prueba no busca evaluar a los/as encuestados/as ni a la institución. En este contexto, te solicitamos que seas lo más honesto/a posible con tus respuestas. De antemano, agradecemos tu valiosa colaboración.

1) ¿Te gusta leer?

Nada ☐ Poco ☐ Regular ☐ Bastante ☐ Muchísimo ☐

2) Generalmente, ¿para qué utilizas internet?

Buscar información de temas generales ☐

Buscar información sobre temas de mi carrera ☐

Entrar en redes sociales, chatear, etc. ☐

Jugar, bajar música, películas, series, etc. ☐

Participar en foros de discusión ☐

Leer literatura ☐

3) ¿Cuántas horas al día pasas en internet?

1 a 3 horas ☐

4 a 6 horas ☐

7 a 9 horas ☐

Más de 10 horas ☐

4) De los siguientes tipos de textos ¿cuál/es de ellos lees por placer? Puedes seleccionar varias opciones.

Textos académicos (manuales, artículos científicos, libros y capítulos teóricos, tesis, etc.) ☐
Textos literarios (novelas, cuentos, poesía, etc.) ☐
Textos de la prensa nacional e internacional (en papel y/o digital) ☐
Textos misceláneos (revistas, ejemplos: Cosmopolita, Muy interesante, Men's Health, etc.) ☐
Textos divulgativos en internet y redes sociales ☐

5) En promedio, cuántos libros (en formato físico o digital) leíste por placer durante los últimos 12 meses. Excluye los textos obligatorios de tu carrera.

Ningún libro ☐
De 1 a 5 libros ☐
De 6 a 10 libros ☐
Más de 10 libros ☐

6) Menciona los últimos tres libros que leíste por placer en los últimos 12 meses.

a._____

b._____

c._____

7) Menciona tus tres libros y autores favoritos

a._____

b._____

c._____

8) En promedio, ¿cuántas horas a la semana dedicas a leer por placer?

Ninguna ☐

De 1 a 5 horas ☐

De 6 a 10 horas ☐

Más de 10 horas ☐

9) ¿Qué porcentaje de los textos obligatorios de tu carrera leíste efectivamente el año/semestre pasado?

25% ☐

50% ☐

80% ☐

100% ☐

10) En promedio, ¿cuántas horas a la semana le dedicas a la lectura de los textos obligatorios de tu carrera?

De 1 a 3 horas ☐

De 4 a 7 horas ☐

De 8 a 10 horas ☐

Más de 10 horas ☐

¡Muchas gracias!

Diccionario de léxico disponible de estudiantes universitarios de Educación Básica y Letras Hispánicas