Universidad Autónoma de Querétaro

Facultad de Psicología

CORPUS DE SUSTANTIVOS MÁS FRECUENTES EN TEXTOS ESCRITOS PARA NIÑOS MEXICANOS, EN MOMENTOS INICIALES DE LA ALFABETIZACIÓN

Tesis

Que como parte de los requisitos para obtener el grado de

Maestra en Desarrollo y Aprendizajes Escolares

Presenta

Alejandra García Aldeco

Santiago de Querétaro

Diciembre 2013

Universidad Autónoma de Querétaro
Facultad de Psicología
Maestría en Desarrollo y Aprendizajes Escolares

CORPUS DE SUSTANTIVOS MÁS FRECUENTES EN TEXTOS ESCRITOS PARA NIÑOS MEXICANOS EN MOMENTOS INICIALES DE LA ALFABETIZACIÓN

TESIS

Que como parte de los requisitos para obtener el grado de

Maestra en Desarrollo y Aprendizajes Escolares

Presenta:

Alejandra García Aldeco

Dirigido por:

Dra. Mónica Alvarado Castellanos

SINODALES

Dra. Mónica Alvarado Castellanos
Presidente — Firma

Dra. Sofia Vernon Carter
Secretario — Firma

Dra. Karina Hess Zimmermann
Vocal — Firma

Dra. Pamela Garbus
Suplente — Firma

Mtra. Norma Fernández Ortega
Suplente — Firma

MDH. Jaime Eleazar Rivas Medina
Director de la Facultad de Psicología

Dr. Irineo Torres Pacheco
Director de Investigación y Posgrado

Centro Universitario
Querétaro, Qro
Diciembre 2013
México

RESUMEN

El objetivo de la presente investigación fue averiguar cuáles y cómo son los sustantivos escritos (estructura y contextos silábicos) más frecuentes presentes en textos de alta divulgación a las que los niños recién alfabetizados están expuestos. El corpus de trabajo estuvo integrado por 360 sustantivos comunes extraídos de 64 textos dirigidos a niños. Con ayuda de las técnicas de la lexicografía de corpus y la estadística descriptiva, identificamos el vocabulario fundamental empleado en los textos. Así mismo observamos que las estructuras silábicas predominantes fueron CV y CV. Las letras iniciales más frecuentes fueron P, C, M y S en combinación con las vocales A y O.

Los resultados del presente estudio fueron comparados con diferentes corpora relativos al español. La utilidad de la presente investigación, fue proporcionar datos sobre la frecuencia de las palabras escritas y sus estructuras silábicas a futuros trabajos que requieran metodológicamente de esta información.

(**Palabras clave**: Sustantivos escritos, análisis de corpus, contexto o estructura silábica).

SUMMARY (ABSTRACT)

The objective of the present research was to discover which are and characteristics of the written nouns (structural and syllabic context) most frequently present in high circulation text to which recently literate children are exposed. The body of the work was made up of 360 common nouns extracted from 64 texts directed at children. Supported on **lexicography corpus technics** and descriptive statistics, we identified the fundamental vocabulary of the written texts. At the same time, there were a predominance of CV and CVC syllabic structures onto noun words. The most frequently used initial letters were P, C, M and S in combination with the vowels A and O.

The results of the present study were compared with different related corpora in Spanish. The purpose of the present research was to provide data about the frequency of written words and their syllabic structure for future work that require **methodologically** of this information.

(**Key Words:** Written nouns, analysis of corpus, syllabic context or structure).

A todas las personas que han contribuido

en mi desarrollo como ser humano

AGRADECIMIENTOS

Este trabajo no hubiera sido posible sin el apoyo de un equipo de personas que a lo largo de mi vida han procurado mi desarrollo. Estas líneas no alcanzarían para mencionar a todos aquellos que me enseñan y acompañan en el día con día. Así que, aunque mencione a algunas personas, hago extensivo el agradecimiento a todos los maestros diarios que por fortuna me rodean.

Agradezco a la Universidad Autónoma de Querétaro porque a través de la Dirección de Investigación y Posgrado contamos con el apoyo del fondo para el fortalecimiento de la investigación (FOFI), recurso que nos permitió formar un equipo de trabajo sólido y contar con el equipo necesario para llevar a cabo esta tesis.

Agradezco también a los excelentes docentes que forman parte de esta casa de estudios, a la Dra. Mónica Alvarado, quien a través de compartir su experiencia como investigadora me transmite el gusto por mirar una y otra vez los datos y buscar la ruta adecuada para interpretarlos. También agradezco profundamente la colaboración en este trabajo, su labor durante mi proceso de formación y su acompañamiento humano a Karina Hess y a Pamela Garbus.

Gracias también, a la colaboración de Cristina Arteaga, Ana Forzán, Valentina Uribe y al equipo de investigación, que con su tiempo y trabajo contribuyeron a que los datos presentados fueran posibles.

Agradezco a la familia García Aldeco por escucharme, apoyarme y enseñarme que a través de la educación uno puede servir mejor a otros. A Juan y a mis amigos, por motivarme y acompañarme en cumplir un reto personal más.

Por último, agradezco a todos los niños con los que he tenido el gusto de compartir un salón de clases por mostrarme lo fascinante que es enseñar y aprender.

TABLA DE CONTENIDO

íNDICE DE TABLAS

ÍNDICE DE FIGURAS

Introducción

El objetivo de la presente investigación es averiguar cuáles y cómo son los sustantivos escritos (estructura y contextos silábicos) más frecuentes presentes en textos de alta divulgación a los que los niños mexicanos recién alfabetizados están expuestos.

Por sus características, este trabajo no plantea un problema de investigación relacionado directamente con la psicología educativa, sino otro ligado con la instrumentación de recursos para el estudio de dicha área, en particular con lo referente al impacto que la exposición a vocabulario escrito específico pudiera tener en los aprendizajes, principalmente ligados con el proceso de alfabetización, que ocurren en los primeros años de la educación primaria.

La presente tesis forma parte de un proyecto mayor en el que, a partir de los recursos de la lexicografía, se pretende hacer una descripción completa del vocabulario escrito presente en los libros de lectura, de los dos primeros años de primaria, distribuidos por la Secretaría de Educación Pública de nuestro país. Como lo mencionamos en un principio, en el estudio que aquí se reporta, abordamos solamente la descripción de los sustantivos.

Este trabajo requirió del empleo de los recursos de la lexicografía computacional y de la lingüística descriptiva. Asimismo, nos dio ocasión de reflexionar sobre qué es una palabra gráfica y su relación con el quehacer de los lexicólogos y lexicógrafos. Así en el primer capítulo, además de mostrar las dificultades en la definición histórica de las

palabras, realizamos una revisión histórica alrededor de la construcción de los diccionarios como vehículos para la descripción de las lenguas.

El segundo capítulo está dedicado a la presentación de un estado del arte sobre el estudio científico del léxico. Se revisan en él los diferentes criterios y recursos de los que el trabajo lexicográfico se ha valido y se presenta la justificación a las decisiones metodológicas que tomamos en la realización de la presente tesis.

En el tercer capítulo se presenta la descripción pormenorizada de la metodología que seguimos en la elaboración de nuestro estudio, para el establecimiento del corpus general y para el análisis del corpus específico de sustantivos.

Presentamos el análisis y la discusión de los resultados en el cuarto capítulo. Además de incluir el análisis del corpus general y el específico sobre sustantivos, presentamos también la descripción silábica del vocabulario fundamental de sustantivos, así como el análisis comparativo entre los resultados de nuestro corpus y otras corpora mayores que describen al español.

Finalmente, el último capítulo de esta tesis está dedicado a presentar las conclusiones de nuestro estudio. En ese apartado señalamos también algunas repercusiones que pudieran tener cabida a partir de nuestra experiencia.

CAPÍTULO I

Antecedentes Sobre el Estudio Lexicográfico

Esta investigación indaga las palabras escritas a las que los niños mexicanos en momentos iniciales de alfabetización están expuestos. Usamos el término "palabra escrita" para referirnos a lo que otros autores consideran como "palabra ortográfica". Al respecto, Zamudio (1999) señala que estas palabras son secuencias de letras separadas por un espacio en blanco que caracterizan a la escritura convencional de una lengua[1]. Cabe señalar que es nuestro propósito establecer un listado de las palabras escritas más frecuentes en los textos de alta divulgación dirigidos al público infantil (alumnos que cursan el primer ciclo de la educación básica).

Indagar sobre las palabras escritas a las que los niños están expuestos es un problema de índole metodológico o instrumental. Nuestro trabajo pretende lograr una base de datos que sirva como recurso para la toma de decisiones en el diseño de trabajos psicolingüísticos que facilite, por ejemplo, evaluar los ítems empleados al solicitar a los niños que lean o escriban determinadas palabras. Como lo mostraremos más adelante, la unidad "palabra" resulta ser particularmente útil para operacionalizar las posibilidades infantiles para emplear la lengua ya sea oral (como en los inventarios para evaluar el desarrollo lingüístico infantil) o escrita.

En la elaboración de esta tesis se consideraron las aportaciones de dos áreas de conocimiento: la lingüística y la psicología. Desde la psicología justificamos la

[1] En este sentido, autores como Smith (1995) definen a la "palabra escrita" como algo con un espacio en blanco del lado que sea.

necesidad de nuestro estudio, ya que las diferentes posturas psicolingüísticas que tratan de dar cuenta de los procesos, sobre todo de alfabetización inicial, requieren, en ocasiones, referencias respecto a los estímulos escritos a los que los niños están expuestos.[2] Más adelante, especificaremos algunos estudios que podrían requerir del corpus que esta investigación genere y la trascendencia, sobre todo, para realizar trabajos en torno a la alfabetización inicial en nuestro país. De la lingüística retomamos la metodología de trabajo que se ha empleado para realizar estudios lexicográficos, los criterios para determinar las fuentes para lograr bases de datos y, sobre todo, los parámetros para la elaboración y clasificación de corpus léxico, en particular, los aportes de la lexicografía del corpus y de la lexicografía computacional.

Si bien especificaremos más adelante los procedimientos lexicográficos empleados para nuestro propósito, nos parece interesante abordar ahora la dificultad para definir una "palabra escrita" de manera que se aprecie con mayor objetividad la unidad de análisis de los lexicólogos.

[2] Podemos referir los trabajos cognoscitivos de Treiman (1993), Chall (1983), Ehri (1991, 1997) y Cuetos (2010) sobre el reconocimiento de palabras escritas en momentos iniciales de la alfabetización. Respecto de la escritura, encontramos los trabajos psicogenéticos encabezados por Ferreiro (1979, 1982), Ferreiro y Teberosky (1979), Teberosky (1990), Vernon (1989), Alvarado (1997).

La Definición Histórica de las Palabras Escritas

La identificación de palabras escritas, como hoy en día las conocemos, ha sido el resultado de un largo proceso de reflexiones lingüísticas y metalingüísticas a lo largo de la historia de la humanidad que han sido motivadas por las necesidades que la escritura y el registro de eventos han generado.

Cabe señalar que los estudios lexicográficos son tan antiguos como la aparición de los llamados "gramáticos" (personas que se especializaban en identificar los vocablos propios de una lengua) ya sea para escribirla, copiarla, traducirla y censurarla en su uso, en contextos orales o escritos.

La historia de la lexicografía está ligada, sobre todo, con la necesidad de interpretar y traducir textos. Podemos ver a través de ella las dificultades que la formalización de palabras (identificación ortográfica y límites acústicos) ha representado. De acuerdo con Zamudio (2010), las palabras gráficamente delimitadas no siempre poblaron las escrituras alfabéticas. Comenzaron a existir muchos siglos después de la invención del alfabeto ya en la era cristiana, porque al parecer los lectores de la antigüedad griega y latina no tuvieron necesidad de ellas.

La palabra en la antigüedad grecolatina

En la cultura grecolatina los textos estaban constituidos por letras que se sucedían de forma regular. El texto se desplegaba de manera continua en columnas paralelas de 15 a 30 letras de ancho y de 25 a 45 líneas de largo formando “páginas” a lo largo del papiro que podían empezar o terminar cortando la expresión en cualquier parte (Martín, 1994, en Zamudio, 2010). La posibilidad de segmentar el texto estaba

entonces determinada por el lector. En ésta época los autores no ejecutaban la escritura de sus obras sino que las dictaban en latín (lengua de la cultura en esa época) a uno o más escribas (Millares, 1975, en Dávalos, 2008).

Los escribas realizaban su tarea en *scriptio continua,* una forma de escritura sin separación entre palabras o frases y sin recursos para la organización gráfica de página. La lectura de este tipo de escritura se basó principalmente en la vocalización, contraria a la lectura a simple vista o en silencio (actividad inusual en la cultura de ese momento). De acuerdo con Parkes (1992), la lectura de un texto escrito en *scriptio continua* requería de varias operaciones. La primera hace referencia a una preparación inicial para la lectura donde el lector debía pronunciar el texto y establecer al menos algunas sílabas que le permitieran aislar porciones de oralidad con significado: *lectio o prelectio*. Esta preparación implicaba una lectura reiterada de la obra de manera que permitiera restituir el significado en su totalidad. Por lo tanto, los lectores requerían de experiencia y de conocimiento previo de los contenidos expuestos (Blecua, 1984 en Dávalos, 2008).

La segunda operación necesaria para leer era interpretar el texto en público, cuidando que la acentuación otorgara sentido al texto: *pronuntiato.* La lectura de los textos estaba dominada por la retórica y la prosodia, por lo que articular de manera adecuada el sentido y el ritmo constituía el ideal de orador. Una vez que se daba lectura en voz alta al texto, se procedía con la *enarratio* que constituía el estudio del vocabulario en su forma retórica y literaria. Por último, se emitía un comentario o *explanatio* sobre el contenido del texto.

Las operaciones realizadas por el lector al tratar de interpretar un texto nos muestran la complejidad de este proceso. Leer sin alguna de estas operaciones, pero en particular sin la *prelectio*, se consideraba un acto presuntuoso propio de ignorantes (Desbordes, 1995 en Zamudio, 2010). La preparación que requerían los lectores no era fortuita; se trataba de una época donde la escritura alfabética se subordinaba a la oralidad y en la cual la *scriptura continua* reflejaba, en un espacio aglutinado, un flujo sonoro ininterrumpido. En este contexto las unidades autónomas de significado no existían gráficamente en los textos, por lo que era imposible identificarlas a la vista. El reconocimiento de palabras debía hacerse mediante el oído.

La palabra durante la Edad Media

Al caer el imperio romano, durante los inicios de la Edad Media, la cristiandad incorporó en su liturgia una nueva forma de vocalización, la *rumminatio,* que consistía en la lectura en voz baja de los textos y que resultaba útil para la reflexión y la memorización de los mismos (Díaz, 2010). Surgieron nuevas necesidades con respecto a los textos; la lectura silenciosa surgió junto con la aparición de recursos gráficos que facilitaran la interpretación y copia de los mismos. Los monjes copistas, quienes se especializaban en la reproducción más o menos estandarizada, incorporaron los recursos gráficos entonces desarrollados a textos prioritariamente sagrados. Fue en este contexto donde se privilegió la lectura silenciosa, tanto por parte de algunos

clérigos lectores a quienes iban dirigidas las reproducciones escritas, como por los monjes copistas.

En el siglo VI se hizo posible la intermediación del texto sin recurrir a la voz. Como lo señala Parkes (1992), la escritura se volvió un medio para transmitir información a través del ojo en lugar del oído, lo que trajo como consecuencia que en la baja Edad Media los lectores comenzaran a leer para ellos mismos. Este cambio posibilitó la lectura en silencio, que llegó a convertirse en el modo de lectura predilecto para los escasos lectores.

La *scripto continua* siguió siendo la forma de escritura cuando los textos se escribían en la lengua de sus usuarios (escribas, copistas o lectores), ya que, al dominar la lengua representada por escrito la interpretación era más o menos sencilla. Sin embargo, en las tierras habitadas por hablantes de lenguas célticas y germánicas, donde la lengua nativa no tenía relación con el latín (lengua en que se escribían la mayoría de los textos, tanto religiosos como académicos), surgieron las primeras segmentaciones de textos y los primeros sistemas de distinción gráfica como posible solución a la dificultad que la copia, edición, traducción e interpretación de textos suponía (Zamudio, 2010).

Las prácticas de lectura y escritura que caracterizaron a la antigüedad no pudieron sostenerse a lo largo de todo el medioevo. Las contribuciones de los monjes copistas en el siglo VII fueron consideradas de acuerdo con (Saenger, 1997, en Dávalos, 2008). La evolución de la comunicación escrita en Occidente Para ellos, el latín era una lengua extranjera aprendida y fundamentalmente literaria. De hecho, el

carácter distante del lenguaje contenido en los textos sagrados les permitió a los monjes reconocer en lo escrito un sistema diferente a la lengua oral por lo que consideraron a la escritura como un lenguaje puramente visual. Esta concepción del lenguaje escrito permitió transformar a la escritura de ser un signo de signos a ser un signo lingüístico con derecho propio (Zamudio, 2010). La palabra escrita se convirtió en un medio para transmitir ideas en vez de ser solo un registro de la palabra hablada.

A finales del siglo VII los monjes insulares, irlandeses e ingleses, experimentaron con el espacio en los textos (acomodo del texto en página) con el propósito de crear apoyos eficientes para la interpretación de la escritura. Parkes (1992) señala que, a consecuencia de haber hallado las gramáticas tradicionales del latín, los copistas irlandeses abandonaron la *scriptio continua* y adoptaron para la interpretación de textos latinos criterios de análisis de las declinaciones y conjugaciones propias de esa lengua, así como las relaciones de los elementos dentro de las oraciones (concordancia). Siguiendo las segmentaciones que las gramáticas presentaban, los monjes lectores pudieron realizar con mayor facilidad el análisis de los textos y la identificación de unidades en la lengua latina. Una vez que estos elementos del discurso fueron identificados, los monjes insulares introdujeron espacios en blanco entre las unidades resultantes. Como consecuencia, se pudo hacer un reconocimiento global de unidades completas de significado (frases).

El proceso de segmentación entre las palabras fue muy largo. La palabra escrita se fijó en los textos latinos después de un análisis que duró un poco más de cuatro siglos. Durante este tiempo se utilizaron diferentes modos de espaciado. Al respecto Zamudio (2004, 2010) explica cinco diferentes criterios, algunos de ellos relacionados

con el patrón de acentuación y otros considerando criterios puramente visuales, como se señala a continuación:

1) Agrupamientos azarosos de letras. Es el tipo más primitivo de espaciado o aireado que durante el medioevo consistía en insertar espacios al interior de una línea sin considerar unidades de significado o incluso de sílabas. La inserción de espacios en blanco se realizaba de acuerdo con criterios gráficos sin ningún parámetro relacionado con la oralidad.

2) Agrupamiento silábico de letras. La segmentación e identificación de las sílabas se realizaba durante la vocalización de los textos. Cuando una línea estaba visualmente muy saturada se introducía un espacio en blanco atendiendo a la parte correspondiente de una sílaba.

3) Agrupamiento jerárquico de letras. En este tipo de separación se utilizaban espacios mayores (salto de línea) para aislar secuencias equivalentes a pequeñas frases. Al interior de estas unidades se introducían espacios en blanco para destacar unidades menores (cláusulas) de manera un tanto aleatoria.

4) Agrupamiento jerárquico de palabras. Este criterio consistía en introducir espacios mayores para aislar una palabra o frase, por lo general con una longitud de 15 o 20 letras y espacios menores o *interpunctus* para distinguir las sílabas dentro de ellas.

5) Bloques sólidos de palabras. Refiere a una variedad de secuencias con un tamaño semejante al grupo anterior pero sin separaciones interiores.

Es importante resaltar que no todos los tipos de espaciado tuvieron la misma aceptación. Saenger (1997, en Zamudio, 2010) destaca que mientras que en Irlanda e Inglaterra una separación muy cercana a la canónica fue usada antes del siglo X, en el continente europeo los agrupamientos largos prevalecieron hasta el siglo XI. Parkes (1992), por su parte, menciona que los agrupamientos gráficos característicos en la Europa de habla romance fueron aquellos que tomaron como límite unidades gramaticales relacionadas con patrones rítmicos que resultaron en agrupamientos coincidentes con unidades prosódicas, primordialmente en la enunciación de los textos. Por ejemplo: *sermodomini* (palabra del Señor), *cumnecessesit* (con necesidad).

La distribución geográfica de los patrones gráficos varió de acuerdo con la relación que los lectores de las islas, actualmente coincidentes con el Reino Unido, (Irlanda, Britania y Escocia) y los del continente europeo tenían con la lengua latina. En España, la separación de palabras siguió un curso diferente al de Italia y Francia por la influencia del árabe que, al ser una lengua consonántica, demandó el empleo de separaciones que facilitaran la división entre términos fonológicamente similares pero semántica y sintácticamente diferentes en posición contigua (Saenger, 1997, en Zamudio 2010).

Los pueblos insulares, al traducir los Evangelios del latín a sus lenguas nativas, pusieron por escrito el inglés y el irlandés hacia finales del siglo VIII, a diferencia de los asentados en el área del antiguo Imperio Romano que comenzaron casi un siglo después (Zamudio, 2010). Como lo señala Saenger (en Zamudio, 2010), la escritura del inglés y el irlandés fue de singular importancia en la historia de la segmentación de

las palabras porque permitió que el latín y las lenguas vernáculas se retroalimentaran y se fijaran criterios comunes de ortografía entre ellas.

Entre los siglos XI y XII se estableció un uso más estandarizado de la separación entre palabras (“separación canónica”) en el que la significación gramatical y léxica se impuso como criterio de segmentación (Zamudio, 2004). Esta segmentación permitió identificar los elementos gramaticalmente funcionales pero sin significado léxico que coincidieron con los monosílabos que actualmente conocemos como palabras funcionales: artículos, conjunciones, preposiciones y pronombres. Así, a pesar de haberse puesto a prueba diversos parámetros relacionados con la vocalización de lo escrito (sílabas y patrones prosódicos), el significado terminó por imponerse como criterio definitivo. Las palabras se convirtieron entonces en los nuevos observables que guiaron la lectura del discurso escrito (Zamudio, 2010).

Con el surgimiento de las universidades en Europa en el siglo XII, en Francia (Sorbona) y en Italia (Bolonia) ocurrieron cambios en la manera de producir textos escritos. Dentro de las instituciones se establecieron talleres editoriales para reproducir copias de ejemplares originales (*exemplar).* Este trabajo se realizaba bajo la supervisión de autoridades académicas.

Como parte de la difusión del cristianismo y por la búsqueda de mantenerse fiel a la palabra de Dios, se mantuvo la preocupación por realizar la copia correcta de las versiones evangélicas y así se intentó disminuir el riesgo de error en las copias con la figura de un corrector.

Así mismo, en esta época surgieron recursos gráficos que facilitaron la lectura y la consulta de los textos: uso de epígrafes, índices, referencias, abreviaturas y foliación de páginas (Le Goff, 1987 en Zamudio 2010. Aparecieron también convenciones textuales como respuesta principalmente a dos necesidades: la de escribir para la clase alta textos normativos de valor jurídico (actas notariales y contratos comerciales) y el interés de algunas organizaciones eclesiales por impulsar la alfabetización entre sus religiosos. Ambas necesidades se presentaron con la mejora de condiciones materiales para la producción de textos escritos: mayor accesibilidad al papel en sustitución de los pergaminos y la invención de la imprenta.

La necesidad de regular la escritura y el surgimiento de la imprenta: el Renacimiento.

La invención de la imprenta, durante el Renacimiento, supuso cambios en el tratamiento y producción de textos escritos. Mucha gente, aunque no supiera leer, estuvo involucrada en la elaboración de textos, por lo que los errores eran comunes a pesar de existir una figura encargada de la revisión: el corrector.

Así mismo, resulta interesante señalar que al inicio de la época renacentista no existía todavía consenso en las formas ortográficas de una lengua. La escritura se modificaba de un taller a otro dependiendo de la trayectoria del corrector asociado. Existió entonces la necesidad de crear tratados y manuales que regularan la ortografía, apareciendo así los primeros diccionarios multilingües y bilingües como el *Lexicón hoc est dictionarium ex sermone latino in hispaniensem* de Antonio de Nebrija (1492), el

Dictionarium latino gallicum (1531) de Robert Estienne y el primer libro para editores elaborado por el italiano Aldo Manuzio: *Manuales y tratados de ortografía dirigida a editores* (1566). Estos textos reflejaron el espíritu renacentista al constituirse como herramientas para la comprensión de una lengua antigua como el latín o el griego e iniciaron una cultura de la lengua normada o estandarizada mediante el sometimiento a las reglas de la gramática.[3]

La Identidad del estado y la normatividad de la lengua: historia de los diccionarios y el quehacer del lexicógrafo

Con el surgimiento de la imprenta sucedieron en Occidente modificaciones sociales con gran trascendencia para la conformación de los Estados modernos: la definición de reinos e imperios y la lucha por la extensión territorial que surgió, sobre todo, a consecuencia del descubrimiento de las Américas. En este contexto político, la delimitación de un reino estuvo también identificada por su lengua, de manera que las lenguas dominantes (de los conquistadores) se extendieron junto con la ganancia de territorios que iban logrando.

La lengua, al considerarse un símbolo importante en la identificación de un reino, motivó la diferenciación entre lenguas muy próximas y el establecimiento de criterios arbitrarios para determinar la soberanía, belleza o pureza de unas sobre otras. Para la

[3] La Gramática de la Lengua Castellana de Elio Antonio de Nebrija (1492) fijó por primera vez una forma gramatical sobre la base ortográfica del español (cánones de representación formal) con el objetivo de asegurar que quedara memoria de la grandeza del imperio español. Así, buscó situar a España en el mismo nivel de prestigio que los imperios de la Antigüedad.

comunidad culta, el trabajo de escribir consistió no sólo en la traducción de textos clásicos, sino también en la determinación de los términos que dieran mejor cuenta de una lengua para que, aunque vernácula, resultara igualmente culta o educada. En este contexto se multiplicó la producción de diccionarios que facilitaran la traducción de términos "difíciles" entre diversas lenguas.

La escritura trascendió a la vida cotidiana, de manera que en ámbitos comerciales hubo también necesidad, ante la falta de traductores, de contar con manuales de equivalencias entre lenguas, por lo que se crearon los primeros vocabularios bilingües de muchas lenguas europeas, americanas, africanas o asiáticas, enfrentadas entre sí. No se trataba de listados de palabras, sino de expresiones o frases completas que se emplearon en diferentes contextos de compra-venta, del tipo *"¿cuánto cuesta…?"* A través de estos documentos escritos, se buscó que la escritura reflejara lo más fielmente posible la interpretación oral del texto: pausas, énfasis y entonación (Sebastián, 2000 en Dávalos, 2008). La concepción de la lengua se modificó. De ser representación verbal de la grandeza del Estado, tomó un carácter más instrumental definido por la necesidad del conocimiento de una parte creciente de la sociedad (Lara, 2007).

Como lo mencionábamos anteriormente, la formación de grandes imperios modernos sirvió para definir un nuevo tipo de diccionario, ya no como una herramienta informativa, sino como un símbolo de identidad. Durante el Renacimiento se reflejó la necesidad de las naciones por reconocerse a sí mismas como culturas distintas a la latina: Inglaterra estableció las bases de su lengua para la posterior expansión en América y la India. Por su parte, Francia logró su principal unidad nacional con Enrique

IV e Italia se reconoció con una lengua culta creada por Dante Alighieri 600 años antes. Alemania, estimulada por el papel que desempeñó la traducción de Lutero de la Biblia, comenzó a formar una identidad escrita propia que conllevó a la primera alusión escrita del “alto alemán”. Por su parte, en España, tras la desaparición de los árabes en la península ibérica, comenzó un proceso de unificación nacional encabezado por Isabel de Castilla y Fernando de Aragón. Estos acontecimientos tuvieron por efecto una reflexión sobre la lengua materna de las nuevas naciones que, con una nueva conciencia, comenzaron la construcción de su identidad nacional.

El diccionario recibió su impulso definitivo en el siglo XVI durante la formación de los Estados Nacionales, cuando además de la transmisión de información, los diccionarios constituyeron también un símbolo de identidad. La idea de lengua que se dio durante ese siglo fue normativa, orientada por el esfuerzo de llegar a equiparar las lenguas maternas con el latín de manera que la lengua literaria se convirtió en el símbolo más importante de identidad y en la base para reflexionar los cánones gramaticales y de corrección.

De acuerdo con Lara (2007), la lexicografía monolingüe apareció en Occidente durante el Siglo XVII como consecuencia de tres fenómenos culturales: el desarrollo de las lenguas modernas como requerimiento para el discurso; la búsqueda de la legitimidad cultural; y la reflexión, de origen filosófico, sobre el origen de las lenguas. Parafraseando a Lara (2007), las lenguas se transformaron en instituciones simbólicas cuya expresión literaria reflejó *el esplendor de la lengua del Estado (sig).*

La nomenclatura de los diccionarios y las definiciones se fijaron entonces; no en términos de equivalencias entre lenguas (condición que habían presentado la primera generación de diccionarios, al inicio de la época renacentista), sino como la descripción de la lengua a través de la introducción de citas textuales de literatos conocidos que avalaran los contextos de uso del vocabulario y la inclusión de entradas léxicas. Esta nueva manera de armar los diccionarios dio comienzo a los principios de la lexicografía.

Es importante destacar que el propósito que persiguió la creación de estos diccionarios monolingües no fue dar respuesta a una necesidad "natural" de la comunidad lingüística por conocer su lengua, sino dar testimonio de la pureza y la autoridad de la lengua a través de los textos correspondientes con los diferentes ámbitos del Estado: político, filosófico, heroico y literario (Lara, 2007).

De acuerdo con Seco (1987, en Lara, 2007) el primer diccionario elaborado con este propósito apareció en 1611: *Tesoro de la Lengua Castellana o Española,* de Sebastián Cobarruvias (también ortografiado como *Covarrubias*). Esta obra, a través del énfasis que daba a la etimología de las palabras, presentó, de manera ligada, la reflexión ontológica de la lengua con su significado "verdadero" (contextualizado en la aparición de los términos dentro de textos literarios). Bajo este esquema, la etimología ayudó a transformar la importancia del diccionario: no sólo proporcionó información sobre el significado de un término y su origen etimológico, sino que con esta información realzaba el valor simbólico y social de la lengua, ya que demostraba que ésta tenía un origen noble (el latín o el griego), además de una historia (tan importante como la identidad de un Reino).

Un año después, en Italia, se publicó el *Vocabolario degli Academici della Crusca*. Este texto respondió a la concepción de lengua que gestó al humanismo: la exhibición lexicográfica de la perfección del italiano (Hausmann, 1989 en Lara, 2007). Los vocablos que conformaron la obra fueron seleccionados de escritos clásicos elaborados entre los siglos XIV y XV, por lo que se consideró un diccionario con contenido de autoridad digno de ser imitado. *El Vocabolario* no se concentró en la etimología de las palabras, sino en la elaboración de definiciones amplias y documentadas, con esta metodología se retomó la tradición lexicográfica desarrollada para los diccionarios multilingües. Por ello, se consideró que este diccionario, en particular, provenía de un método lexicográfico riguroso y preciso Pfister, 1989 (en Lara, 2007).

El diccionario de la *Crusca (1612)* impulsó la actividad purista sobre la lengua y fue ejemplo a seguir para dos Estados contemporáneos: Francia y España. En Francia, Jean Chapelain, autor del primer proyecto del diccionario, planeó rescatar fragmentos de escritores ya fallecidos que ejemplificaran su obra (imitando el método de Italia) y propuso la recreación de contextos de la oralidad cuando fuera imposible encontrar ejemplos escritos de los vocablos de una lengua. Los vocablos, de acuerdo con el autor, debían escribirse seguidos de una marca gráfica que mostrara que su aprobación se otorgaba por su uso común. De esta manera se dio inicio a la incorporación de vocablos de la oralidad dentro de los diccionarios.

La Academia Francesa, creada por mandato del cardenal Richelieu, asumió en el año de 1634 la responsabilidad de definir a la lengua francesa y buscó, por medio de su diccionario, que los usuarios de la lengua incorporaran el vocabulario *honnêtes*

gens[4]*,* empleado por los poetas y oradores representativos de la época. La lengua de los *honnêtes gens* fue considerada como evidencia de todo aquello que podía servir para la nobleza y elegancia del discurso y que complementaba las cualidades estéticas que un hombre podía tener en la vida civil (Popelar, 1976 en Lara, 1997).

A lo largo del S.XVIII la incorporación de fragmentos escritos fue desapareciendo y se fortaleció la idea de contar con autoridades, en el dominio de la lengua, que pudieran definir y ejemplificar los vocablos contenidos en los diccionarios. De esta manera, las Academias decidieron eliminar las citas textuales de fragmentos escritos y confiar a una sola persona la responsabilidad de crear un diccionario que posteriormente debía someterse al juicio de la compañía.

En España, con la conciencia de que el español había alcanzado su esplendor durante el Siglo de Oro, se buscó fijar el funcionamiento y riqueza de la lengua española a través del establecimiento explícito de reglas que la protegieran de la corrupción. Es por esta razón que por iniciativa de Juan Manuel Pacheco, con la aprobación de Felipe V, el 3 de octubre de 1714 se fundó la Real Academia Española (en adelante RAE), institución que bajo el lema *"Limpia, fija y da esplendor",* buscó establecer las voces y vocablos de la lengua castellana con su mayor elegancia y propiedad[5].

A diferencia de la Academia Francesa, la Española se concentró en la incorporación de la documentación histórica de la lengua. No sólo presentaba el origen

[4] La denominación "Honnêtes gens" se puede traducir, del francés, como "gente honrada" y hace referencia a la gente común de un lugar.
[5] La fundación de la RAE siguió la misma tendencia de la Academia de la Lengua Francesa fundada en 1697.

etimológico de las palabras, sino que también mantuvo la incorporación de textos escritos que mostraran las diferencias que el uso de un vocablo había presentado en diferentes textos o periodos de la escritura. De esta manera, se incorporó a la práctica lexicográfica la documentación histórica de la lengua, constituyendo esto el inicio del principio filológico que caracteriza a muchos diccionarios contemporáneos.

Para 1726, la RAE creó el primer Diccionario de la Lengua Castellana, conocido como *Diccionario de Autoridades.* La novedad de este documento fue que en él se expusieron los usos de la puntuación, producto de contar con una base de fragmentos escritos suficientes que posibilitaba, incluso la normalización de estos elementos propios de lo escrito, y a la separación gráfica de las palabras (entendidas ya como unidades con significado). Cabe señalar, como lo expresa Zamudio (2004), que estas definiciones se han mantenido hasta nuestros días [6] .

La creación del *Diccionario de Autoridades* provocó cambios sociales importantes. En España se creó una nueva clase social *nobilitas literaria* favorecida con privilegios del rey, exenciones de pago y ayudas económicas para asistir a funciones públicas, que permitieran poner en práctica sus ideas para el bien de la monarquía.

Junto con el surgimiento del Diccionario de Autoridades, durante el Siglo XVII, la construcción del diccionario y su uso en los países continentales permitió la reflexión de la lengua literaria. El estudio de la lengua común quedó fuera de su ámbito; los

[6] En 1713 se publicó el primer tomo del diccionario de Autoridades. Trece años después se publicó el sexto y último tomo de la obra. Tras 26 años de labor conjunta, unido en un solo tomo y sin ejemplos, este trabajo ha tenido veintiún ediciones hasta la última versión elaborada en formato electrónico en 2001 (Lara, 2006).

diccionarios, que hasta antes del *Diccionario de Autoridades* habían servido como catálogos simbólicos representativos de la calidad del vocabulario, comenzaron a verse como obras de consulta general (Lara, 2007).

El diccionario en Inglaterra

La transformación del propósito del diccionario no se dio de manera homogénea en todos los países. La historia de los diccionarios en Inglaterra fue diferente a la que se dio en los estados continentales; a pesar de que también se comenzó a formar un Estado nacional, la lengua inglesa quedó delegada a segundo término debido a la acción latinizante de los eruditos de la corte isabelina (Robertson y Cassidy, 1954 en Lara, 2007).

El primer diccionario monolingüe elaborado en lengua inglesa fue la publicación de Robert Cawdrey (1604), *A Table Spphabeticall*. Este diccionario constituyó un buen precedente de la lexicografía inglesa, ya que indicaba las lenguas (hebreo, francés, griego) de las que provenía el inglés y decía enseñar la verdadera escritura y la comprensión de las palabras usuales del inglés que son difíciles. Esta publicación tuvo poca relevancia estatal debido a que el 90% de las palabras contenidas habían sido publicadas ya en un libro de gramática y léxico creado en 1596, llamado *Edmund Cootes English Schoole Master.* Con el mismo propósito, más tarde se creó *An English Expositour: Teaching the Interpretation of the Hardest* de John Bullokar (1616).

De acuerdo con Noyes (1943, en Lara 2007) existió una corriente postergada en la lexicografía monolingüe inglesa: la de los maestros que, en 1582, veían como

necesidad urgente la creación de un diccionario que edificara el inglés y facilitara el uso de la lengua materna. Esta inquietud fue ignorada por los lexicógrafos, quienes se centraron en los elementos más excéntricos y menos permanentes de la lengua, razón por la cual, de acuerdo con Noyes, 1943 (en Lara 2007) la aparición del diccionario inglés se retrasó por más de un siglo.

Años más tarde, la creación del *Universal Etymological English Dictionary* (1721) de Nathaniel Bailey inauguró la presencia de los diccionarios monolingües completos en inglés. Este diccionario, al igual que el resto de los diccionarios monolingües, se interesó por la etimología y se centró en su uso como herramienta para explicar las palabras difíciles o *inkoenterms* presentes en la literatura de la época. Mientras se multiplicaban los diccionarios para esclarecer el significado y uso de términos difíciles o poco frecuentes, fue creciendo la necesidad de ligar la lengua con la identidad de los Estados; de manera que, al igual que los estados del centro de Europa, comenzaron a usar los diccionarios para fijar el vocabulario que demostrara el prestigio de la lengua oficial. El proceso fue igual al que se dio en los Estados Continentales pero con una variante significativa: la necesidad de difundir los diccionarios a una mayor parte de la sociedad. Ya no solamente la Corte y los eruditos se interesaban por tratar la lengua sino también un nuevo público creciente: “the public school”.

Si bien dentro de un sector de ingleses educados o *públic cultivé* (eruditos, nobles y burgueses) surgió la propuesta de crear una academia de la lengua al estilo europeo, las condiciones sociales en la Inglaterra del siglo XVIII la imposibilitaron; en primer lugar, por la falta de personas que se propusieran aptas para crearla y, en segundo, porque el soberano no mostró interés por instituirla. En cambio, se afirmó el

poder instrumental de los diccionarios y otros portadores textuales definido por la necesidad de conocimiento de una parte mayor de la sociedad.

El papel de la burguesía ejemplifica lo determinante que fue la presión social en esta transformación. Gracias a estos nuevos hombres ilustrados, los periódicos dejaron de ser portadores exclusivos de noticias mercantiles para convertirse en un medio de debate y reflexión. La función de la lengua fue servir como instrumento de conocimiento social haciendo que la labor de los lexicógrafos no fuera solo servir al Estado, sino también proporcionar parámetros lingüísticos a las agrupaciones burguesas dedicadas a la charla y a la formación de lo que más tarde sería "la opinión pública".

La lexicografía vino a definir las nuevas dimensiones burguesas y la nueva concepción de la lengua con la publicación en 1755 de Samuel Johnson del *Dictionary of the English Language.* Esta edición, a pesar de las dudas de su propio autor sobre la autoridad que portaba y de que perseguía los mismos ideales de fijación y pureza que otras publicaciones, elaboradas en Europa en siglos anteriores, fue recibida con entusiasmo por el público en general. Las publicaciones de prensa y la gente de esa época coincidieron en que el diccionario suplió la necesidad de una academia de las bellas letras ya que no negó la tradición literaria y la importancia del Estado hasta ahora constituida e incluyó valores del hombre ilustrado, como la razón y el igualitarismo (Lara, 2007).

De acuerdo con Robertson y Cassidy (1954, en Lara, 2007), la autoridad de los diccionarios surgió de dos fuerzas presentes en los siglos XVII y XVIII: una artística (literaria) y otra social (arribo de la clase media a la prominencia social). Ambas, junto

con la necesidad de corrección en el uso de la lengua, provocaron que el diccionario tuviera valor por sí mismo y se convirtiera en una herramienta pedagógica que puso a disposición una lengua correcta y aceptada por la misma sociedad. Los lexicógrafos de esta época lograron que tanto los círculos letrados del Estado como el público general otorgaran al vocabulario presente en los diccionarios un valor de verdad (en tanto que eran representativos de la lengua) y con ello se volvieron un parámetro referencial para la comunicación social.

La idea del diccionario como suprema autoridad lingüística Wells, 1973 en Lara, 2007) permitió que, por primera vez en la historia de la lexicografía monolingüe, se utilizara incluso como referencia en cuestiones de jurisprudencia en1806. El diccionario se extendió más allá de los círculos letrados del Estado y se situó como garantía de comunicación social y como marco de referencia para el discurso público.

Con la necesidad de uso del diccionario en diferentes contextos lingüísticos (áreas de conocimiento y dialectos) los diccionarios sufrieron modificaciones en cuanto a la especificidad lingüística que atendían. De esta manera, se convirtieron en herramientas para la descripción de la lengua empleada por un grupo o subgrupo social particular. Por su relevancia social, su elaboración requirió la intervención de una nueva ciencia: la lexicología, ciencia que a través de la definición del término “palabra” y su contexto se encargó de estudiar (junto con su estructura, composición y variación) el vocabulario de una lengua. Así mismo, surgió la formación de profesionales centrados en la observación científica del comportamiento y uso del vocabulario de una comunidad lingüística: los lexicólogos.

La lexicografía en Estados Unidos

Si bien los estudios lexicográficos más comunes están relacionados con la elaboración de los diccionarios, también han servido para realizar la caracterización de minorías lingüísticas. A continuación describimos cómo la lexicografía contribuyó a la creación de una identidad nacional en los Estados Unidos al ser una nación recién independizada de Gran Bretaña.

La lexicografía estadounidense se desarrolló como respuesta a la necesidad de legitimar una nación. Reconoció la labor de la lexicografía inglesa de perfeccionar el inglés durante un siglo, pero reservó para los Estados Unidos, "país de luces y libertad", la labor de llevar al inglés a su mayor perfección (Lara, 2007). Buscó que el inglés se constituyera como lengua nacional de forma que sustituyera la pluralidad de las lenguas maternas existente entre los estados.

Con más claridad que como aconteció en Inglaterra, la burguesía angloamericana no dio lugar a la institución de una academia que vigilara la pureza y la propiedad de la lengua inglesa, sino que, nuevamente, un diccionario se constituyó en la norma: *The American Spelling Book* (1828). Con esta edición Noah Webster, maestro de escuela y autor del diccionario, respondió a una búsqueda interna de regularidades en la propia lengua. Concebía al diccionario como una obra para el lector común y como una guía para el correcto uso de la lengua vernácula, por lo que buscó reglas internas en el lenguaje que permitieran guiar al alumno en el aprendizaje de la lectura y la escritura. Después de reflexionar sobre la ortografía inglesa y su uso, llegó

a la conclusión de que la analogía, la costumbre y el hábito son los mejores parámetros para regular el uso de las palabras. Estos nuevos criterios modificaron el quehacer lexicológico. Como lo señala Campos (2007) (en Lara, 2007), Webster ofreció, por medio de la razón, una idea científica de la fijación de una lengua y su ortografía (en un diccionario), idea que pronto se convirtió en una ley general.

The American Spelling Book, contó con 400 ediciones y 15 millones de ejemplares y fue muy bien recibido por el público gracias a que contenía las normas propias del inglés americano frente a las británicas. Asimismo, se constituyó en una mercancía al salir del patronato del Estado y pasar a las casas editoriales. Los efectos de su nuevo carácter comercial aparecieron en la segunda mitad del siglo XVIII. Dado el impacto social de este material, aunado a las nuevas tendencias en la elaboración de diccionarios especializados, los diccionarios entonces editados incluían vocablos técnicos (no literarios: científicos, sociales y regionales).

El uso de diccionarios se socializó muy rápidamente de manera que la producción de estos libros de consulta también se vio incrementada así como las ganancias económicas, que hicieron que algunas casas editoriales se dedicaran exclusivamente a su elaboración y tiraje. En este contexto se inició la lexicografía enciclopedista (Lara, 2007).

La rivalidad entre lexicólogos ingleses y británicos los llevó a interesarse por la variedad regional angloamericana creando diccionarios que dieran lugar a la lengua nacional. La consideración de variantes dialectales contribuyó a poner en duda la idea de la lengua literaria única, formada por los humanistas del siglo XVI.

Durante la segunda mitad del siglo XVIII y el siglo XIX la lexicografía privada se desarrolló a través de un proceso permanente de copia y reproducción de los diccionarios académicos, ejemplares que ganaron un lugar como acervo de la sociedad. A mediados del siglo XIX se llevó a cabo en Estados Unidos la "guerra de los diccionarios", momento histórico en el que las casas editoriales se enriquecieron por medio del plagio del corpus de los diccionarios originales.

La reproducción masiva de diccionarios, junto con la necesidad de añadir en ellos voces de reciente creación, pronto dejó entrever la insuficiencia de los diccionarios académicos. Por esta razón, no solo en América, sino también en Europa, el diccionario se convirtió en blanco de ataques lexicográficos individuales. La inclusión de términos que no se acogieran a la pureza normativa impuesta por las academias se justificó con la creación de una nueva ciencia: la filología. Ésta no se opuso al carácter normativo de la lengua, pero sí mostró interés por documentar su evolución mediante la recuperación de datos empíricos (Anglada y Bargalló, 2007).

La influencia de la filología en el quehacer lexicológico

En los inicios del siglo XIX, el estudio de la historia comenzó a requerirse en la elaboración de los diccionarios a partir de una nueva corriente de pensamiento: el romanticismo alemán. Esta corriente, interesada por recuperar la tradición folclórica del pueblo, impulsó la creación de la filología a través del análisis histórico de los textos escritos. Se creía que de esta manera se podría tener acceso al espíritu de los pueblos del pasado y, con ello, poder apreciar las manifestaciones actuales.

La nueva ciencia definió un nuevo rumbo para la lexicografía monolingüe. Los diccionarios anteriores, conocidos como diccionarios de *autoridades,* podían verse como históricos, al ser registros de momentos anteriores de una lengua y, en un trayecto histórico de diccionarios, recuperar la evolución de la misma. Dado que en diferentes momentos los diccionarios no fueron suficientemente actualizados, o, por su necesidad normativa, registraron más el "deber ser" lingüístico que la descripción de la realización de la lengua, los filólogos tuvieron que hacer estudios, a partir de la producción literaria de diferentes épocas, que les posibilitara constatar la evolución real de la una lengua.

El primer diccionario histórico europeo fue financiado por la casa editorial Weidman de Leipzig: *Deutsches Wörterbuch* de Jacob y Wilhelm Grimm (1852), ejemplar que tuvo como objetivo documentar y estudiar el nuevo alto alemán, desde el siglo XV hasta su época, a partir de fuentes literarias. Para la inclusión de vocablos en esta obra se investigó el origen de los mismos de manera rigurosa e histórica, con bases documentales, por lo que la legitimidad simbólica del trabajo fue dada, ya no por el Estado, sino por el pueblo mismo. Sin embargo, la editorial no concluyó este proyecto; la elaboración de este diccionario se delegó a diversas instituciones hasta su culminación en 1971, por lo que su publicación resultó anacrónica (Lara, 2007).

Si bien el trabajo filológico involucró a algunos lexicólogos, otros continuaron el trabajo relacionado con la elaboración de diccionarios especializados. De hecho, muchas de las academias de la lengua se mantuvieron en la elaboración del trabajo como lo habían venido realizando e incluso, como en el caso de la Academia Española, se opusieron a la ideología historicista en la elaboración de diccionarios, no sólo por el

rastreo que tendrían que hacer de los vocablos, sino también por considerar que no podrían dar apertura a vocablos vulgares (tendencia de los diccionarios históricos a incluir los vocablos del uso habitual presentes en los textos escritos de diferentes épocas). Así, en la undécima edición del diccionario de la Academia Española se señala que: “se mantenía firme en su decisión de no sancionar más palabras nuevas que las indispensables, de recta formación incorporadas al castellano por el uso de personas doctas”.

Para este punto de la historia eran tres los tipos de diccionarios: los *normativos y generales* (en cuanto a ser representativos de la lengua oficial) con censura de expertos; los *especializados,* que describían los vocablos de grupos y subgrupos de hablantes; y los *históricos,* que realizaron tanto filólogos y lexicólogos independientes como instituciones universitarias. Es así que entre los años 1853 y 1855 salió a la luz el primer diccionario enciclopédico en castellano, *Diccionario enciclopédico de la lengua española,* realizado por Gaspar y Roig.

La visión del diccionario como un inventario histórico modificó la labor del lexicógrafo, quien comenzó a buscar fuentes literarias para seleccionar y rastrear vocablos que describir en sus diccionarios. Una de las acciones representativas de este quehacer fue la fundación, en 1842, de la Philological Society de Londres, institución encargada de estudiar la historia de la lengua inglesa. El ideal de esta institución fue realizar un estudio acucioso de la lengua que permitiera incluir, dentro del *New English Dictionary* (que en 1915 se convirtiera en el *Oxford English Dictionary*), todas las palabras de uso de esa lengua, independientemente de ser consideradas doctas o

populares, y sin tomar en cuenta los juicios de los lexicólogos encargados de este gran proyecto.

Las críticas expuestas por Richard Chenevix Trench (fundador de la Philological Society de Londres) en su obra *On Some Deficiencies in Our English Dictionaries* (1857), ayudó a perfilar la función moderna de los diccionarios que fueron concebidos con un sentido práctico, con bases filológicas. La materialización de esta visión se concretó 71 años después de ser propuesta con el *New English Dictionary*. Este diccionario definía significados y daba la etimología de los vocablos, la pronunciación, la ortografía y el uso, teniendo como fuente principal el análisis filológico de las obras literarias (doctas y cotidianas) inglesas más representativas del siglo XIX.

A finales del siglo XIX la filología dio paso al desarrollo de la lingüística como disciplina científica interesada en el análisis de las lenguas. Desde esta nueva disciplina el quehacer de los filólogos y lexicólogos comenzó a diferenciarse.

Los parámetros de la lingüística descriptiva no coincidieron con la filología que inspiró los diccionarios enciclopédicos. El caso más representativo de esta incompatibilidad fue la crítica hacia la tercera edición del diccionario de Noah Webster en Estados Unidos: *Websters Third New International Dictionary* (en adelante 3W). La inclusión de vocablos poco aceptados por los conservadores y puristas de la sociedad angloamericana desataron las críticas de periodistas, quienes tomaron los “mismos ejemplos” para mostrar cómo el diccionario era promotor de la propuesta “speaks as you go” (“habla como quieras”) contrario a los valores educativos y culturales de la

nación. Es importante destacar que las críticas se realizaron tomando ejemplos aislados sin considerar al texto mismo.

Phillip. B. Gove, director de la tercera versión del diccionario, justificó su metodología para la inclusión de vocablos mediante el criterio de *uso*. Este criterio fue objetado por Mario Pei, divulgador de la lingüística moderna en Estados Unidos, quien argumentó que si el parámetro de inclusión era el uso, habría que considerar: ¿el uso de quién? Como respuesta a este cuestionamiento surgió un nuevo parámetro como parte de la metodología lexicográfica, la frecuencia (Lara, 2007).

La crítica que tuvo el 3W fue la consecuencia del enfrentamiento entre un valor científico de la lingüística y una tradición autoritaria heredada del desarrollo de la lexicografía europea desde el siglo XVI. A partir de entonces los valores de la ciencia dejaron de debatirse públicamente, como sucedió con la burguesía ilustrada, y comenzaron a imponerse desde los círculos científicos al margen de la sociedad.

De acuerdo con Lara (2007) la lingüística moderna, con su carácter científico, tuvo que decidir entre dos posturas: por un lado, la necesidad social de contar con una idea de la lengua; por el otro, comprender cómo se gesta ésta y evoluciona en la sociedad. De acuerdo con Lara (2007), la lingüística debe tomar en cuenta la dimensión social de la lengua ya que puede colaborar con el establecimiento de mejores métodos y técnicas lexicográficas que permitan considerar con objetividad la riqueza de las manifestaciones verbales existentes. El diccionario monolingüe se convirtió así en una herramienta importante para la lingüística, porque permite observar la idea que se tiene sobre la lengua y su conocimiento social como fenómeno colectivo.

En suma, desde la aparición del diccionario hasta hoy en día, los objetivos y métodos lexicográficos, y en consecuencia el “hecho diccionario”, se ha construido en relación con la evolución histórica de los países. Estados Unidos se constituye como un parteaguas en la metodología lexicográfica. Como se vio, con el ánimo de forjar su identidad nacional, los estadounidenses incluyeron las variables de uso y frecuencia que dan forma al diccionario actual, el cual se abordará en el siguiente capítulo.

CAPÍTULO 2

EL ESTUDIO CIENTÍFICO DEL LÉXICO

En el capítulo anterior abordamos la historia de la palabra escrita; pudimos observar que aunque la aparición de la lengua es ontogenéticamente primaria no fue sino hasta la aparición de sus manifestaciones escritas que el hombre pudo comenzar a reflexionar sobre ella. A través de los manuales de lengua, y más tarde por medio de la creación de los diccionarios, el ser humano cobró conciencia de que la lengua vale no por ella misma sino por la manera en que nos permite concebir al mundo y actuar sobre él. También discutimos cómo el desarrollo de la lexicología en el siglo XIX posibilitó la creación de diccionarios concebidos como catálogos verdaderos de la comunidad lingüística. Su función semántica y su normatividad los convirtieron en objetos verbales notables dignos de reflexión (Lara, 2006). Junto con el desarrollo de diccionarios, la lexicografía, ciencia que se encarga de su elaboración, evolucionó. El objetivo del presente capítulo es mostrar las aportaciones que la evolución de la lexicografía ofreció a la lingüística y a la psicolingüística para el estudio científico del lexicón[7] principalmente a partir del trabajo con corpora.

El estudio científico del léxico a través de corpora

Los lexicógrafos, personas especializadas que guían el proceso de elaboración de diccionarios, se han encargado de estudiar y documentar cuáles son los elementos necesarios para que un diccionario pueda considerarse representativo de la lengua o variante dialectal que intenta describir. Algunos aspectos a considerarse son:

[7] El término “lexicón” refriere a una realidad colectiva que comparten los hablantes de una lengua y a la capacidad que tiene un hablante de internalizar su vocabulario (Luque, 2004).

1) a quién va dirigido el diccionario (usuario)
2) el corpus a partir del cual se realizará el análisis de datos o vocablos
3) los ejemplos que se ofrecerán en la edición para ilustrar los contextos de uso de los vocablos
4) las ilustraciones, en caso de necesidad, que acompañarán a los vocablos

Para el presente trabajo resulta importante retomar las reflexiones que se han generado en torno al corpus (sobre el que se elaboran los diccionarios) porque nos interesa mostrar cómo la posibilidad de tener bases de datos reales influyó en el trabajo psicolingüístico a propósito de contar con la identificación de vocablos prototípicos en los diferentes momentos de la adquisición de la lengua.

En este capítulo comenzaremos por mostrar qué entendemos por el término “corpus”, luego indagaremos la clasificación que existe para las corpora y cuáles son sus criterios de elaboración. Después, intentaremos explicar cuáles son sus usos y funciones, sobre todo para los estudios psicolingüísticos.

Definición de *Corpus*

El término *corpus* se ha usado en lingüística con dos acepciones. La primera, hace referencia a la recopilación de material lingüístico, ya sea de textos orales o escritos, bajo un propósito de investigación. Esta acepción se ha utilizado en estudios de lingüística aplicada, como en el aprendizaje o adquisición de lenguas. Por ejemplo, cuando se reúnen producciones, orales o escritas, de aprendices en diferentes contextos (sociales, de edad, etcétera). La segunda acepción de *corpus* designa a un conjunto extenso de textos seleccionados para representar a una lengua. La lingüística de corpus es considerada como un área científica, dependiente de la lingüística teórica, que busca representar el vocabulario natural de una lengua. La lingüística de corpus, al igual que la lingüística general, estudia las características de una lengua: su estructura, la manera en cómo se habla, su historia, su origen y los fenómenos humanos que se manifiestan en ella. Se ha especializado en el trabajo con datos reales que permiten reproducir con máxima fidelidad las características del objeto de estudio, por medio del estudio de las palabras en su contexto. De ahí que la base de su estudio se finque en un corpus amplio y representativo de la lengua objeto de estudio.

El interés que perseguimos en la presente investigación es caracterizar a las palabras escritas a través de la elaboración de un corpus restringido que nos posibilite analizar, al menos, los sustantivos prototípicos presentes en textos literarios dirigidos a niños que cursan los primeros años de la educación primaria. Para ello resultó pertinente apegarnos a las definiciones y métodos que propone la lingüística de corpus. De acuerdo con Sinclair (2004) (en Torruella y J. Listerri, 1999 p.7), uno de los especialistas modernos en lexicología de corpus, un corpus se define como:

> *A collection of pieces of language that are selected and ordered according to explicit criteria in order to be used as a sample of the language.*

La definición anterior se enriquece con la propuesta de Crystal (1991, p.32, en Parodi, 2008, p.102) que señala:

> *Una colección de datos lingüísticos, ya sea de textos escritos o de transcripciones de habla grabada, que pueden ser utilizados como punto de partida para descripciones lingüísticas o como un medio de verificación de hipótesis acerca de una lengua.*

En una tercera definición, Pérez (2002) señala que es un conjunto de textos seleccionados bajo un criterio lingüístico específico que representa una muestra del uso que los hablantes nativos hacen de una lengua.

Estas tres acepciones del término “corpus” nos permitieron considerar que un corpus es una herramienta para describir el lenguaje que requiere ser seleccionado bajo criterios lingüísticos específicos.

En su concepción, el estudio con corpora (lingüística de corpus) es muy antiguo, pero sus alcances y metodología se han modificado a lo largo del tiempo. A continuación desarrollamos brevemente la historia de la lingüística de corpus de manera que se aprecie cómo es que las corpora han cobrado importancia como fuente de información para realizar estudios lingüísticos.

Evolución Histórica de la Elaboración de Corpus

La historia de la lingüística de corpus documenta que, a la par que comenzó a surgir la "guerra de los diccionarios", surgió el interés por crear listas léxicas y listas de frecuencias que constituyeran, sobre todo, una herramienta para la enseñanza de las lenguas. Es así, como a finales del siglo XIX, surgió el texto *Häufigkeitswörterbuch der deutschen Sprache*, trabajo que fue publicado de forma privada por Käding en Berlín, en 1897-1898 (McEnery & Wilson, 1996). Käding, con la ayuda de varios miles de colaboradores, procesó un corpus de doce millones de formas aproximadamente, incluidas en textos alemanes. Este texto se creó con la finalidad de ayudar a mejorar la taquigrafía de la época y su enseñanza, al mostrar cuáles eran las combinaciones de letras y sílabas más frecuentes.

A partir de 1920, el interés por la creación de listas de frecuencias, léxicas o gramaticales, se acrecentó, por lo que este tipo de trabajos se extendió a lenguas muy diversas. Dentro de las primeras listas de palabras frecuentes destacó la labor realizada por Thorndike en 1921, con la publicación de la primera lista de palabras más frecuentes del inglés. Esta lista, fue constituida a partir de 41 fuentes textuales y se conformó por un total aproximado de 4,5 millones de palabras. Este trabajo fue considerado un corpus relevante porque a partir de él el mismo Thorndike realizó el primer diccionario dirigido al ámbito escolar: *The Teacher Word Book*. Este diccionario fue considerado el primero dirigido al público infantil y combinó los principios de la psicología con los del aprendizaje. Fue una edición dirigida a maestros y estudiantes y su intención primordial fue mostrar el vocabulario que un buen estudiante "debía" saber.

Años más tarde, en 1944, Thorndike colaboró con Lorge para analizar un corpus de 18 millones de formas y producir una lista de las 30,000 palabras más frecuentes del inglés, misma que influyó de manera significativa en el contenido de los textos escolares publicados en años posteriores.

A finales de la década de 1950, se originó el proyecto *Survey of English Usage de Randolph Quick* cuyos datos constituyeron la base de la primera gramática del inglés titulada *A Comprehensive Grammar of the English Language*. Este trabajo, aunque no utilizó datos informatizados, sirvió de base para la posterior compilación informática de datos (Martín, 2009). Aunque no existe documentación sobre el proceso de elaboración de esta gramática, sabemos que este trabajo tomó la mitad de los datos de la lengua oral y el resto de la lengua escrita.

El avance de la tecnología posibilitó que en 1961 surgiera el *Brown University Standard Corpus of Present-Day American English* creado por Francis y Kucera. Este trabajo es considerado una obra pionera de la disciplina, debido a que fue el primer corpus concebido para ser electrónico, por lo que la metodología utilizada para su elaboración estableció las características de las primeras corpora textuales construidas con ayuda del análisis computacional.

El *Brown Corpus* estuvo compuesto por un millón de palabras seleccionadas en 500 textos americanos distribuidos en 15 categorías textuales distintas. Dado que el tamaño de la base de datos no fue grande por las limitaciones de las máquinas de la época, la representatividad de este trabajo se consiguió al reunir un conjunto relativamente amplio de muestras textuales de tamaño reducido. Aunque este trabajo

fue muy popular en Europa, su repercusión en Estados Unidos fue escasa (Martín, 2009).

Los trabajos lexicográficos en español han sido menos que los realizados para el inglés. Resaltan los trabajos publicados por Keniston (1937) que describieron tanto las estructuras sintácticas de la prosa española del s. XVI, como la del s. XX.

Por su parte, Juilland y Chang (1964) crearon el *Frecuency Dictionary of Spanish Words* como resultado del análisis de un corpus de aproximadamente medio millón de formas. A partir del trabajo con tipos diferentes de aparición de una palabra (types) y número de ocurrencias (tokens), este trabajo mostró la lista de los 5,024 lemas[8] más 'frecuentes' del español (según el conjunto de factores utilizados por los autores: frecuencia, dispersión y uso) y las formas asociadas a ellos y dio cuenta de la variedad morfológica del español.

Aunque los estudios realizados en contextos hispanoparlantes comenzaron a tomar fuerza a partir de 1964, no fue sino hasta 1975 que se realizó el primer diccionario basado en un corpus moderno, el *Diccionario del Español de México*. Este diccionario fue realizado por Luis Fernando Lara con una muestra de más de 2 millones de palabras que sirvieron de base para la posterior publicación en 1982 del *Diccionario Fundamental de México*.

En España, desde una perspectiva psicolingüística, Alameda y Cuetos (1995), caracterizaron al español a través de la elaboración de un subcorpus compuesto por

[8] Se entiende por lema a la forma o entrada con la que se identifica genéricamente un conjunto de vocablos relativos. Por ejemplo, niñas, niñitos, niña estarían bajo la cabeza lemática "abuelo". En español existe la convención de expresar un lema en la forma masculina y singular de los sustantivos.

dos millones de palabras distribuidas en diferentes portadores textuales: 25% periódicos, 15% ensayos literarios, 10% revistas científicas y 50% novelas. Este trabajo contiene listas de formas, sílabas y combinaciones de letras. Es un subcorpus que no está lematizado y ha servido de referencia, aún hoy en día, para trabajos de investigación relacionados con el reconocimiento visual de palabras.

Ya en nuestro siglo, en el año 2000, salió a la luz el corpus *LEXESP* creado por Sebastían, Martí, Carreiras y Cuetos. *LEXESP* es un corpus conformado por 5,029, 930 palabras obtenidas de textos escritos entre 1978 y 1995. Para su elaboración, se consideró la inclusión de tres diferentes tipos textuales: 40% novelas, 30% periódicos y 30% ensayos periodísticos. Este corpus sirvió como base de datos para que en el mismo año Alameda y Cuetos crearan el *Diccionario de Frecuencias de las Unidades Lingüísticas del Español,* publicación que mostró las palabras más usuales en el español.

Con un propósito diferente a la recopilación y descripción de información, Davies, creó en 2006 un corpus destinado a facilitar el aprendizaje del español como segunda lengua, *Corpus L2,* el cual estuvo construido sobre un conjunto de veinte millones de formas aproximadamente y contuvo únicamente los 5,000 lemas más frecuentes (sin indicación de las formas asociadas), con los índices de frecuencia correspondientes.

Respecto a la producción de corpora recientes, podemos referirnos al corpus *Subtlex-Esp* (2011) y al *Spanish-Word Frecuencies* (2012). *Subtlex* es un corpus formado a partir de subtítulos de series y películas circulando en internet. Este trabajo

consta de un total de 39,935,628 palabras; de estas 1,222,111 son palabras en español y el resto de origen inglés. Dentro de la metodología para su elaboración, se reportó la exclusión de no palabras, símbolos, nombres propios, onomatopeyas, palabras derivadas, flexionadas y verboides.

Por su parte, el *Spanish-Word Frecuencies* (2012) es un instrumento basado en un corpus de 41 millones de palabras tomadas de series televisivas transmitidas entre los años 1950 y 2009. Además de lograr la identificación del habla prototípica empleada para la televisión, muestra, particularmente, las modificaciones dialectales de las diferentes generaciones de hablantes representados en la televisión.

Gracias a la sistematización electrónica, en la actualidad el trabajo con corpora considera bases electrónicas de gran alcance por el número de palabras y variedad de contextos de recopilación. Pueden consultarse en la WEB. Una de las limitantes de estos estudios radica en que la mayoría de las corpora se refiere al habla inglesa. Algunas de las más conocidas son las siguientes:

Corpus	Vocablos incluidos
Spoken English Corpus (SEC)	50,000
Louvain Corpus of modern English drama	1’300,000
Lancaster, Oslo & Bergen Corpus (LOB)	1’000,000
Brown Corpus	1’000,000
London-Lund Corpus of Spoken English	500,000
T.O.S.C.A.	1’500,000
Birmingham Collection of English Texts	40’000,000
International Corpus of English (en desarrollo)	

El análisis de corpora ha evolucionado con gran velocidad y ha permitido dar cuenta de aspectos de la lengua que antes no eran observables por ejemplo, la

frecuencia de aparición de algunas palabras, sus regularidades e irregularidades de acuerdo con los diferentes contextos de uso de la lengua (Rojo, 2008).

Como lo hemos mostrado hasta aquí, son muchas las corpora logradas para las diferentes lenguas. En la actualidad, a partir de los análisis de Gries (2006 en Rojo, 2008) existe consenso en identificar cinco principios generales de todas ellas:

a) El análisis ha de basarse en el formato electrónico de los textos para que la recuperación de datos se realice de forma automática.
b) Los vocablos de un corpus tendrán que ser representativos de la lengua o del contexto lingüístico del que fue extraído.
c) El análisis será sistemático y exhaustivo.
d) El análisis utilizará datos estadísticos para su descripción.
e) El análisis se realizará sobre la base de listas de frecuencias, concordancias y colocaciones.

Estos principios van de la mano con los avances tecnológicos en materia de analizadores electrónicos. Con el paso de los años, los lexicógrafos incluso han generado programas computacionales para el procesamiento específico de información lexicológica. Por lo anterior, con la lingüística de corpus también la lingüística computacional se ha desarrollado, como se describe a continuación.

Desarrollo de la Lingüística Computacional

La lingüística computacional se define como el estudio de los sistemas de computación útiles para la generación y comprensión de lenguas naturales. De acuerdo con Grishman (1986 en Faber, Moreno y Pérez, 1999), posee tres funciones principales: la

creación de interfases para la consulta de datos, la traducción automática y la recuperación automática de información a partir de textos en una lengua natural.

La lingüística computacional, al igual que la lingüística de corpus, requiere de metodología especializada que permita analizar a profundidad el léxico de una lengua. Para poder obtener un estudio específico del lenguaje, los especialistas desarrollaron el uso de nuevos métodos que con el paso del tiempo han llegado a considerarse, por algunos, disciplinas en sí mismas[9]. En este estudio fueron significativas al menos dos: la lexicografía de corpus y la lexicografía computacional (Faber, Moreno y Pérez, 1999).

Metodología de la lingüística de corpus / Lexicografía de corpus computacional

La lexicografía computacional regula el uso de medios técnicos computacionales que apoyan los procesos que sigue la elaboración de un diccionario, tanto en su microestructura (almacenamiento de información durante la recopilación de entradas o procesos de etiquetado) como en su macroestructura (ordenación del conjunto de materiales que forman el cuerpo de un diccionario, tipo de palabras, de entradas y orden en que se van a ofrecer). En el presente estudio, el conocimiento sobre la lexicografía computacional nos permitió valorar las herramientas de los distintos analizadores léxicos existentes y nos dio la posibilidad de elegir la más pertinente de acuerdo con la finalidad de nuestro estudio.

[9] Aún no existe un consenso entre los lingüistas sobre considerar a la lexicografía como una ciencia en sí misma. Lara (2006), considera que no es una ciencia, sino una metodología que define y enseña los métodos que se siguen para la elaboración de diccionarios.

Existen diversos programas que posibilitan el análisis lingüístico y que son denominados analizadores léxicos. Las herramientas básicas que deben tener para que se consideren completos son: frecuencia de aparición de palabras, concordancias, lematización, detección de unidades recurrentes (colocación) y análisis morfológico (tagging) (Pérez, 1999). A continuación, desglosamos qué implica cada función:

a) Frecuencia de aparición.- Es la posibilidad que tiene un software de contabilizar el número de veces que se encuentra un término. Los analizadores crean listados considerando dos datos: palabras/formas (types/ tokens), es decir, el número total de palabras de un texto frente al número de palabras diferentes que aparecen en el mismo. Estos listados son de gran importancia para la lexicografía ya que ayudan a decidir las voces o vocablos que se deben incluir en un diccionario.

b) Concordancia.- Esta función permite observar la concordancia de una palabra en una lista (también conocida como KWIC Key Word in Context). Genera un índice de todos los ejemplos que existen en un texto determinado (o en un grupo de textos) de la palabra que previamente se selecciona.

c) Lematización.- Esta función permite asignar diferentes formas de una palabra a una misma forma canónica. Por ejemplo, agrupar formas como “abuela”, “abuelita”, “abuelo”, “abuelos” bajo el lema “abuelo”. En el español, la forma canónica de los vocablos se expresa, en el caso de los sustantivos, en singular masculino.

d) Colocación.- Es una herramienta que permite mostrar discurso repetido en un corpus a partir de la solicitud de una frase por parte del investigador. El analizador léxico muestra en una pantalla las palabras que anteceden (con un número de caracteres

predeterminado) y suceden a la frase solicitada. Esta función pretende ayudar a identificar tendencias y estructuras. A partir de una tabla de datos muestra los contextos en los que una palabra (previamente seleccionada) aparece. Muestra a las palabras por columnas y contabiliza las veces que la palabra de la primera columna aparece junto a la palabra de la segunda y en qué posición se encuentra (una, dos, tres palabras a la derecha, a la izquierda, etc.).

e) Análisis morfológico.- Permite caracterizar a una palabra de acuerdo con su categoría gramatical, forma y clase, dependiendo de su uso en una oración**.**

Los analizadores léxicos se han desarrollado rápidamente en la última década. En la actualidad existen programas que además de realizar un análisis de las palabras gráficas permiten contabilizar el tiempo de reacción de un sujeto al enfrentarse a ellas. Por ejemplo, en el estudio *Subtlex* (2011), se observó el reconocimiento de palabras a partir de la lectura de una lista de pseudopalabras. Para ello, se utilizó el programa *DMDX, Forster & Forster (2003).* Esta herramienta permitió registrar la latencia de respuesta al reconocer una palabra ya que permite sincronizar el tiempo de reconocimiento visual con el reconocimiento auditivo. El uso de este programa también se reportó dentro de la metodología de elaboración del corpus LEXESP de la Universidad de Barcelona (2000).

Gracias a su metodología y a los avances tecnológicos, la lingüística de corpus ha aportado a los estudios lexicográficos la posibilidad de alcanzar un *total accountability* Leech, 1992, 1995 (en Rojo, 2008) es decir, dar cuenta del comportamiento que los elementos estudiados muestran en todos los contextos reales posibles. Gracias a esta posibilidad, las corpora hoy en día ofrecen a la lingüística

datos que, mediante un análisis manual, no alcanzarían la misma profundidad (Cabré, 2007). Es por eso que tienen un lugar privilegiado entre los estudios lingüísticos y poseen incluso publicaciones especializadas a nivel internacional.

Parámetros lexicográficos para la construcción de un corpus

El uso de criterios específicos para la elaboración de un corpus crea una diferencia entre el corpus y un conjunto de archivos de cualquier índole (enciclopedia, biblioteca, colección de revistas, grabaciones de voz de un contexto determinado), por lo que la metodología utilizada para su creación adquiere relevancia. En este sentido, el diseño del contenido y el modo de representación de un corpus es sumamente importante debido a que debe buscarse su replicabilidad, por el esfuerzo económico y humano que supone (Torruella y Listerri, 1999).

Los parámetros que determinan la elaboración de un corpus no son iguales en todos los manuales de lexicografía pero sí similares. Distintos autores como Kennedy (1997), Sinclair (2005), McEnry y Xiao (2005) y Ávila (2006) concuerdan en que para la construcción de un corpus se deben cuidar los siguientes aspectos: contenido, muestra, representatividad, variedad, cantidad, homogeneidad, objetividad, validez y confiabilidad. A continuación, desglosamos a qué se refiere cada uno.

a) Contenido y muestra. Este criterio hace referencia a que el contenido del corpus debe ser creado según criterios externos (la función comunicativa de los textos del corpus) y no criterios internos (los referidos a la lengua de los textos) (Sinclair, 2005). La muestra debe seleccionarse a partir del objetivo de estudio e incluso, de ser

posible, debe consistir en documentos o transcripciones de eventos del habla completos aunque difieran en tamaño.

b) Representatividad y variedad. Debido a que la población léxica es ilimitada, resulta imposible determinar la totalidad de palabras a partir de un listado. El criterio de representatividad delimita el contenido de un corpus. Este criterio especifica que un corpus debe ser una muestra representativa de la lengua, de modo que los resultados de su análisis sean válidos y generalizables.

Se entiende por representatividad a la capacidad que tiene el corpus para mostrar el léxico que pretende medir, independientemente del número de individuos que lo utilicen o de la frecuencia que en ese momento histórico tenga el vocabulario estudiado. De acuerdo con Sinclair (2005), la representatividad, al igual que el contenido y la muestra, serán dependientes de la finalidad con la que se construya el corpus. Este criterio está relacionado con el muestreo del corpus, que puede ser de dos tipos: *longitudinal* o *transversal* ('cross sectional'). Sinclair considera dentro del parámetro de representatividad al "principio de equilibrio". Al utilizar este término el autor refiere que el corpus debe estar equilibrado y contener muestras representativas de todo tipo de lengua (oral y escrita), sobre todo para los llamados *monitor corpora*.

Por su parte, Lara (2006) propone que para lograr que un corpus sea representativo en la elaboración de la base de datos se seleccionen textos de diversas fuentes en cantidad suficiente. Para este lexicógrafo es importante que al elaborar un corpus se considere la amplia variedad de textos que la riqueza de la lengua posee: textos literarios, periodísticos, informativos, transcripciones de conversaciones en diversos contextos entre otros. Autores como Hopkins , Hopkins y Glass (1997), entre

otros, enfatizan este criterio dentro de la metodología de corpus. Para ellos, en investigación es más importante la representatividad de la muestra que la preocupación por el tamaño de la misma. Ávila (2006) concuerda en la relevancia que tiene este criterio para construir un corpus al argumentar que de la representatividad de la muestra de un estudio dependerá su validez y confiabilidad.

c) Cantidad. Otro de los criterios lingüísticos que se cuidan al elaborar un corpus es la cantidad de palabras que lo conforman. El tamaño que debe tener un corpus para considerarse confiable depende de su riqueza léxica (tipos de palabras, types) más que del número de ocurrencias (tokens) (Torruella y Listerri, 1999).

La cantidad de vocablos que debe contener un corpus varía dependiendo de la finalidad que persiga y de la tipología de corpus que se construya. Por ejemplo, de acuerdo con Tribble (1997), un corpus especializado requiere una cantidad menor de palabras que si pretendiera dar cuenta de una lengua en su totalidad, siendo un tamaño aceptable 25,000 palabras.

Para Lara (2006) la cantidad de palabras incluidas en un corpus es suficiente en la medida que las ocurrencias dentro de los vocablos crecen (tokens) y los tipos de palabras disminuyen (types). Rojo (2008), por su parte, menciona que al elaborar un corpus no es suficiente tener una gran cantidad de textos, puesto que la cantidad sola no es garantía. De acuerdo con este autor, la construcción de un corpus requiere ser definida y organizada para que represente con validez a la lengua.

d) Homogeneidad. De acuerdo con Sinclair (2005), el objetivo de un corpus es alcanzar la homogeneidad de sus componentes (composición y estructura uniformes).

Para ello, en este criterio se verifican dos aspectos. Por un lado, se debe cuidar mantener una cobertura adecuada de textos, por el otro, evitar los textos atípicos (rogue texts). De tal manera, una vez finalizada la etapa de recolección de datos, se debe examinar cada texto para eliminar los que no corresponden con la finalidad del corpus que se elabora.

e) Validez, confiablidad y objetividad. La validez de un instrumento se valora de acuerdo con el grado en que mide lo que pretende medir, Sampieri (2010). De acuerdo con este autor, existen diferentes tipos de evidencia de este concepto:

i) Validez de contenido. Es el grado en que la medición representa al concepto o variable medida. En un corpus, es la medida en la que se representa a la lengua que se estudia. Para que un instrumento sea válido, requiere tener representados todas o la mayoría de las variables de dominio de contenido (variables a medir) (Bohrnstedt, 1976, en Sampieri, 2010). La validez de contenido generalmente se determina a partir de la teoría sobre el tema que se trata y de los estudios antecedentes.
ii) Validez de criterio. Un instrumento muestra su validez al ser comparado con algún criterio externo que pretende medir lo mismo. Si diferentes instrumentos miden conceptos similares sus resultados deben ser similares también (Sampieri, 2010).
iii) Validez de constructo. Es la validez que se otorga a partir del éxito con el que se representa y mide un concepto teórico. Se considera la más relevante de todos los tipos de validez (Grinnell, Williams y Unrau, 2009, en Sampieri, 2010). Cuanto más elaborada se encuentre la teoría que apoya a la hipótesis, la

validación del constructo se acercará más a la validez general de un instrumento de medición (Sampieri, 2010).

iv) Validez de expertos. Algunos autores consideran la opinión de expertos en el tema “voces calificadas” como un criterio de validez para un instrumento. Su opinión califica el grado en el que aparentemente el instrumento mide la variable estudiada (Sampieri, 2010).

La validez de un instrumento se complementa con la confiabilidad del mismo. El término confiabilidad refiere al grado en que la aplicación de un instrumento produce resultados consistentes y coherentes. Este aspecto varía dependiendo del número de ítems con el que cuente el instrumento de medición (Sampieri, 2010).

La validez y la confiabilidad de la medición de una variable dependen de las decisiones que se tomen para operacionalizarla y lograr una adecuada comprensión del concepto, evitando imprecisiones y ambigüedad. En caso contrario, la variable corre el riesgo inherente de ser invalidada debido a que no produce información confiable (Ávila, 2006).

Además de la validez y confiabilidad un instrumento debe ser objetivo. El término objetividad refiere al grado en que un instrumento es susceptible a sesgos y tendencias de los investigadores que lo administran, califican e interpretan. De acuerdo con Sampieri (2010), toda recolección de datos debe cumplir con este criterio, cuidando evitar factores que puedan afectar los datos, como puede ser utilizar instrumentos desarrollados en el extranjero que no han sido validados al contexto en el que serán aplicados.

f) Contraste. Este principio hace referencia a la posibilidad que un corpus tiene para compararse con otros, ya sea por niveles de aprendizaje o entre corpora de fuentes similares. Por ejemplo, es posible la comparación entre corpora de lenguaje de aprendices y corpora de nativohablantes (Prieto, Mosqueira y Vázquez, 2009).

g) Criterios estructurales o de documentación. La elaboración de un corpus es un proceso elaborado que requiere de un trabajo en diferentes fases. Este parámetro refiere a la necesidad de que todos los criterios para determinar la estructura de un corpus sean reducidos en número y claramente separables los unos de los otros. Todo el procedimiento debe estar documentado detalladamente. En lo posible, debe contener información sobre los contenidos y los argumentos que justifican las decisiones tomadas durante el proceso de elaboración para que, si en un momento dado el investigador obtiene resultados 'extraños', se pueda corroborar, por ejemplo, si existió un error de diseño o de selección de textos (Sinclair, 2005).

g) Etiquetado. De acuerdo con este criterio, cualquier información acerca del texto (aparte de la información alfanumérica: palabras y signos de puntuación) debe ser almacenada en un formato aparte del texto puro para después ser fusionada con el texto si la aplicación informática lo requiere (Sinclair, 2005). La creación de archivos separados tiene la finalidad de que el corpus pueda analizarse en su forma inicial sin información adicional.

Los criterios anteriores permiten a los trabajos de lingüística de corpus generalizar los conocimientos que forma. Sin embargo, es importante resaltar que no todos las corpora son iguales. Los criterios de elaboración requerirán de modificaciones dependiendo de la finalidad que el corpus persiga. En lo que sigue, describiremos

algunos de los diferentes tipos de corpora que existen, lo que nos permitirá más adelante justificar la metodología seleccionada para elaborar nuestro corpus.

Tipología de corpora

Existen diferentes tipos de corpora. Autores como Sinclair (1996), Torruella y Llisterri (1999) han propuesto clasificaciones de los distintos tipos de corpora en función de dos criterios: 1) la finalidad para la que fueron creados y 2) los parámetros que se utilizaron para su elaboración (porcentaje y distribución de los textos que lo componen, especificidad de los textos que lo componen, cantidad de textos que se compilan, sistema de representación utilizado para la obtención de base de datos, entre otros). Así, aunque los términos entre autores varían, las características de las corpora son similares.

Dentro de la lexicografía de corpus se expresa la dificultad para determinar dentro de una sola categoría a un corpus, ya que es posible que un trabajo pertenezca a dos o más categorías, dependiendo de su uso. A continuación, desglosamos las categorías de clasificación que serán de utilidad para nuestro trabajo.

a) Clasificación de acuerdo con el sistema representación de la que se obtiene la base de datos

Bajo este criterio las corpora se pueden clasificar en: corpora textuales, orales o mixtos. Las corpora *textuales* constituyen muestras de la lengua escrita en sus diferentes tipos textuales: periódicos, libros, cartas, recetas, etc. Pueden ser textos de cualquier año, área geográfica y variedad lingüística. Podemos encontrar dentro de esta categoría al corpus *IULA*, desarrollado por el Instituto Universitario de Lingüística Aplicada de la

Universidad Pompeu Fabra, Barcelona (Cabré y Bach, 2004). Este corpus, contiene textos escritos en cinco lenguas diferentes (catalán, castellano, inglés, francés y alemán) de las áreas de economía, derecho, medio ambiente, medicina e informática, además de documentos paralelos sobre estas materias.

Las corpora *orales*, por su parte, constituyen muestras de la lengua oral (transcripciones ortográficas de la lengua hablada o grabaciones de conversaciones acompañadas de su transcripción). Tienen como propósito caracterizar, desde un punto de vista lingüístico, a la lengua hablada al observar, entre otras cosas, palabras cortadas, reconstrucción de segmentos omitidos por el hablante, vacilaciones, elementos fáticos (afirmación, duda, interrogación, negación), simultaneidad de turno de palabras y errores de producción del hablante, etc. Las fuentes de donde se obtienen los datos que integran estas corpora son diversas, desde un laboratorio formal de fonética hasta la grabación de un diálogo cotidiano entre dos o más personas. Un ejemplo de este tipo de corpus es *CORLEC, Corpus Oral de Referencia del Español Contemporáneo,* de la Universidad de Madrid. Este trabajo está registrado como el primer corpus de habla espontánea del español. Se realizó entre 1991 y 1992, está conformado por 1,100 palabras tomadas de eventos de habla en debates, eventos deportivos, entrevistas, documentales, noticiarios, anuncios publicitarios y eventos religiosos (Marín, 1992). El *CORLEC* se incorporó a un corpus mucho más grande perteneciente a la misma categoría: el *Subcorpus Oral del Corpus de Referencia del Español Actual CREA.* Este último fue realizado a partir del 10% de textos orales del CREA. Está conformado por 9 millones de registros, aproximadamente, procedentes de

más de 1,600 documentos. Por su tamaño, el subcorpus oral del CREA es relevante para describir al habla hispana (véase www. rae.com).

Por su parte, las corpora *mixtas* son aquellas que utilizan formato oral y formato escrito para alimentar su base de datos. Tal es el caso del corpus generado por César Hernández Alonso dentro del proyecto *EGREHA, Estudio gramatical del español hablado en América*, dirigido por la Universidad de Valladolid. Este corpus incluyó dos fuentes de datos. Por un lado, incorporó los materiales escritos del corpus de la norma culta MC-NC, creado por Lope Blanch en 1964. Por otro, recopiló las transcripciones de entrevistas orales.

b) Clasificación de acuerdo con la especificidad de los textos que lo componen

Bajo este rubro las corpora pueden ser generales o especializadas. Un corpus *general* busca reflejar la lengua en su ámbito más amplio. Para ello, recupera todos los géneros textuales posibles sin límite de cantidad. Por ejemplo, el *BNC British National Corpus* contiene una muestra de la lengua inglesa desde 1975 hasta la actualidad (Oxford, 2013). Por su extensión y variedad, las corpora generales sirven como referencia para la elaboración de diccionarios y para la validación de diversos estudios. Dentro de esta categoría también se encuentra el *Corpus de Referencia del Español Actual (CREA).* Para su elaboración se compilaron textos literarios, periodísticos, científicos y técnicos, así como transcripciones de grabaciones de la lengua oral y de medios de comunicación, correspondientes a los últimos veinticinco años (1975-1999) en el español (Real Academia Española, 2001).

El corpus *especializado* es el que busca recuperar los datos que describen una variante dialectal particular. Esta tipología pretende observar con mayor profundidad ámbitos específicos del lenguaje. Algunos ejemplos son el *Corpus del ámbito médico especializado en oncología OncoTerm, (2002)* y el corpus especializado en el área de puertos y costas *PuertoTerm, (2007).* Ambas corpora realizaron la caracterización del lenguaje específico de su área y crearon una interface de términos comunes, con lo que lograron estandarizar el lenguaje en una población específica: pacientes oncológicos y personal de salud en el primer corpus, y trabajadores en el área de puertos en el segundo.

c) Clasificación por la representatividad cronológica del corpus

Por otro lado, las corpora pueden referir al análisis de vocablos prototípicos de una época concreta. Por ejemplo, el *Corpus del Español de Brigham Young University* (Davies, 2002),[10] es un corpus compuesto por 100 millones de palabras reunidas en textos de los siglos XIII al XX.

d) Otras clasificaciones.

Cuando un corpus refiere a los vocablos empleados por un solo autor o una sola fuente, entra en la categoría de "corpus canónico". Un ejemplo de este tipo de corpus es el *TIME Corpus*, corpus que conjunta el léxico de la revista Time desde 1920 al 2000. Por otro lado, bajo el criterio de la cantidad de texto que se toma de un documento, un corpus puede ser de tres diferentes tipos: corpus de textos completos, corpus de referencia y corpus léxico.

[10] Disponible en http://www.corpusdelespanol.org/

El corpus de *textos completos* (whole corpus)[11] recoge textos completos de los documentos que lo constituyen (cuentos completos, todas las secciones que conforman un periódico, etcétera). Pérez (2002) considera que en la actualidad casi todas las corpora de nueva construcción pertenecen a esta categoría, gracias a la capacidad de las computadoras para almacenar y manejar información. De acuerdo con Sinclair (2005), considerar para el análisis lexicográfico textos completos disminuye los problemas de las diferencias que puede haber dentro de un texto, por lo que el autor considera que el corpus de texto completo es un instrumento apto para la diversidad de estudios lingüísticos posibles. El *Corpus Dinámico del Castellano de Chile (Codicach*), es un ejemplo de corpus de textos completos, ha servido para cateterizar el habla de los chilenos, a través de mostrar cerca de 800 millones de palabras.[12]

El corpus de *referencia* recoge fragmentos de los textos que lo conforman. Busca conformar una muestra representativa de las variedades más importantes de una lengua, así como de sus estructuras y vocabulario general (Pérez, 2002). De acuerdo con Sinclair (2005), en la elaboración de este tipo de corpus debe tenerse especial cuidado con el equilibrio y la representatividad que se obtiene, ya que, como su nombre lo dice, sirve como una base de datos para comparación; incluso, se utiliza para la elaboración de diccionarios. El *CEMC, Corpus del Español Mexicano Contemporáneo* (1921-1974), elaborado por Lara, García y Ham (1980) ejemplifica a este tipo de corpus. Este trabajo brindó las características del español mexicano

[11] Estas mismas características de un corpus se clasifican para Torruella y Listerri, (1999), en el término corpus textuales.

[12] La información detallada de la composición y proceso de elaboración de éste corpus se encuentra disponible en la base de datos electrónica. http://sadowsky.cl/codicach.html.

contemporáneo a través de una colección de contextos de uso de las palabras y sirvió como referencia para la elaboración del Diccionario del Español de México. Otro corpus que ha sido utilizado como referencia para estudios psicolingüísticos es *CHILDES* (Bryan Macwhinney, 1995) este corpus busca mostrar datos de la adquisición del lenguaje, conjuntando transcripciones de audio y video realizadas con niños desde 1960 hasta el día de hoy. Es un corpus relevante porque incluye más de 20 idiomas que permitieron la creación de más de 130 sub-corpus especializados. A partir de 1990, fue computarizado, lo que facilitó el manejo de sus volúmenes. Gracias a la posibilidad de digitalizar la base de datos, hubo un aumento significativo en el número de estudios en torno a la adquisición del lenguaje infantil. Hasta la fecha, este corpus ha sido citado en más de 3,000 estudios (CHILDES, 2003).

El corpus *léxico o simple* recupera fragmentos pequeños de textos. La selección de los mismos es cuidada tanto en la constancia de géneros textuales como en su longitud. Dentro de esta categoría se encuentra el *Corpus del Español Mexicano Contemporáneo,* ya que la selección de sus textos se limitó a 2,000 palabras gráficas por categoría textual.

f) Clasificación por número de palabras contenidas

Las corpora también pueden clasificarse por el número de palabras que contienen. Desde este criterio se distinguen las corpora cerradas y abiertas (monitor). Las corpora *cerradas* están compuestas por un número finito de palabras, que se determina de forma previa a su recopilación. Una característica de estas corpora es que, una vez finalizada la etapa de recopilación de datos, no se añaden más textos al análisis. Esta

característica convierte a las corpora cerradas en una herramienta útil cuando se pretende estudiar estados de la lengua de un momento en particular. *El Corpus del Español Mexicano Contemporáneo CEMC* se encuentra dentro de esta categoría. Está conformado por 996 textos escritos por autores mexicanos desde 1921 hasta 1974.

Las corpora *abiertos o de monitor* se han hecho posibles gracias al tratamiento computacional de la información. La creación de un corpus monitor, comenzó con Clear,1987 (en Pérez, 2002), quien propuso crear un corpus en constante crecimiento, en el que se añadiera de forma periódica y permanente material nuevo, al mismo tiempo que se eliminan cantidades equivalentes de material antiguo. Renovar la información que contiene un corpus persigue el propósito de ofrecer al lingüista la posibilidad de observar cambios recientes en el uso de la lengua.

De acuerdo con Torruella y Listerri (1999), en la actualidad no es necesario establecer un límite al tamaño del corpus, por lo que el criterio propuesto por Clear ,1987 (en Pérez, 2002) ha quedado en desuso. Sin embargo, al añadir información, existe la condición de que siempre crezca con una porción equivalente a los estratos anteriores. Al tener cambios constantes el corpus se considera como dinámico y vivo, características propias de la lengua. El corpus *CREA* es un ejemplo de esta tipología porque pretende ser una referencia del español actual y su base de datos se encuentra en constante renovación.

g) Clasificación por el número de lenguas incluidas

Un corpus se puede clasificar en monolingüe y bilingües o multilingües. Como su nombre lo indica, las corpora monolingües integran vocablos de una sola lengua. El

International Corpus of English ICE es un ejemplo representativo de este tipo, por caracterizar al inglés con textos escritos y traducciones a partir de 1989. Para la elaboración de este corpus, veinticuatro equipos de investigación documentan alrededor del mundo las variantes del inglés, por lo que representa a profundidad las características de esta lengua.

Se denomina corpus *bilingüe* al trabajo que compila textos en dos lenguas diferentes y corpus *multilingüe*, al que incluye tres o más lenguas. Esta tipología de corpora ha sido utilizada en los últimos años con frecuencia dentro de los trabajos de traducción automática. Existe dentro de esta categoría una distinción entre las corpora que presentan textos diferentes en idiomas diferentes y los compuestos por un mismo texto traducido a varios idiomas (corpora paralelas) (Pérez, 2002).

Dentro de esta categoría existe una subcategoría "corpora bilingües o multilingües comparables". Las corpora comparables son aquellas construidas con características similares en distintos idiomas. Distinguimos con estas características al proyecto *NERC Network of European Reference Corpora*, que reúne las lenguas de la Unión Europea[13]. En español, podemos encontrar al *Corpus Kings College* que recoge muestras del español de la Península, de Argentina y de América del Sur.

h) Corpora de variantes dialectales

Existen también corpora que dan cuenta de variantes dialectales específicas. La creación de este tipo de materiales constituyó un parteaguas para la tradición lexicográfica, ya que, si bien los estudios lexicográficos más comunes están detrás de

[13] Para mayor información sobre este corpus consultar el portal: http://www.ilsp.gr/en/infoprojects/meta?view=project&task=show&id=31

la elaboración de los diccionarios, también han servido para realizar la caracterización de variantes dialectales específicas. Tal es el caso del corpus *Varilex* (2003) elaborado por Ruiz, A y H. Ueda de la Universidad de Tokio. Este corpus reúne la variación léxica del español en el mundo a partir del 2003 hasta la fecha. Para la creación de su base de datos, recopila información en cuestionarios electrónicos (vía internet) dirigidos a hablantes de una lengua, mismos que son revisados y publicados por colaboradores en los diferentes países de habla hispana.

El corpus *BASYQUE*, también ejemplifica el estudio de la variación dialectal. Este corpus conformó una base de datos con ejemplos de variables dialectales compilados a partir de tres fuentes de información: cuestionarios, videos y textos literarios. *BASYQUE* apoya la creación de plataformas digitales que facilitan la unificación de un idioma con sus diferentes dialectos (Uria, Hulden, Etxeberria y Alegría, 2011).

En el apartado anterior mostramos los lineamientos establecidos por la lingüística de corpus para la creación de corpora y cómo es que a partir de los criterios que retoma se clasifican. Además de los procedimientos metodológicos propios del trabajo con corpora para la elaboración de esta tesis fue importante considerar las aportaciones de la lexicografía general en el estudio con las palabras como unidad de análisis. A continuación, describimos algunas consideraciones que los lexicógrafos sugieren incluir para su estudio.

La Objetivación de la Palabra como Unidad de Análisis

De acuerdo con Lara (2006) los estudios lingüísticos, por lo general, han prescindido de considerar a "la palabra" como una unidad de análisis, poniendo en duda su existencia como unidad verbal y desestimando su valor unitario. Al intentar dar cuenta de qué clase de sistemas son las lenguas, la unidad palabra se ha considerado como un elemento secundario. Existe una fuerte tradición de realizar análisis lingüísticos considerando a los morfemas o a los fonemas como elementos suficientes para dar cuenta del funcionamiento de la lengua. En otro extremo, estaría también la tendencia de realizar análisis discursivos más amplios en los que las frases podrían ser las unidades de estudio. La definición de palabra resulta problemática aún para aquellos que se especializan en el estudio del lenguaje.

En el diccionario se le define como un conjunto de trazos gráficos continuos separados por espacios en blanco. Sin embargo, esta definición gráfica no es suficiente. Lara (2006) considera que la definición gráfica de la palabra se dejaría afuera a todas aquellas lenguas que son habladas y no escritas. Una segunda razón, es la capacidad metódica que tiene la palabra para poder delimitar una unidad que parece tener valor organizativo y cognoscitivo en todas las lenguas. En su curso de Lexicografía, Lara (2006) emprende el reto de explicar, desde la lingüística, las razones por las cuales la unidad palabra es tan evidente para los hablantes de una lengua[14] y cómo la formación de ésta unidad está determinada por la lengua desde la que se estudie, en este caso el español.

[14] Lara (2006) considera que cuando los hablantes pueden realizar reflexiones tipo "unidad de cita", muestran su sensibilidad hacia la palabra. Por ejemplo, la palabra "abuelo" es la unidad de cita (o lema para el lexicógrafo) de las palabras: abuelito, abuelos, abuelita, abuela. Otro ejemplo de la sensibilidad de los hablantes hacia las palabras, sería la pronominalización.

De acuerdo con este lexicógrafo, existen tres condiciones necesarias para que la unidad palabra exista: sus características fonológicas, características semánticas y características morfológicas. A continuación, definimos el carácter fonológico de la palabra a partir de un procedimiento de la lingüística descriptiva: determinar un concepto a partir de su significado. Describir algunos rasgos de las características fonológicas de las palabras, nos permitirá mostrar características propias del español que más adelante compararemos con los resultados obtenidos del corpus de sustantivos más frecuentes del que nos ocupamos en esta tesis.

Las características silábicas de las palabras del español

La palabra puede ser considerada como una combinación de sílabas que a su vez están integradas por combinaciones de fonemas. Los fonemas son, de acuerdo con Lyons (1981, en Quinteros, 1997), las unidades fonológicas que permiten establecer diferencias de significados y poseen una función distintiva basada en la oposición fónica. Los fonemas al combinarse, ofrecen al signo su forma de expresión (Lara, 2006). Identificar los fonemas y establecer las relaciones estructurales posibles entre ellos constituyen la primera condición para determinar la palabra en cada lengua.

Cada lengua posee su propio sistema fonológico que además de regir los tipos de fonemas posibles, también regula su composición en la sílaba y su organización dentro de la palabra (Serra, Serrat, Solé, Bel y Aparici, 2000). El español está conformado por fonemas vocálicos y consonánticos. Los vocálicos son los fonemas que por sí solos, aisladamente o combinados entre sí, pueden formar palabras o sílabas (Alarcos, 1986). Las combinaciones silábicas de una lengua son más o menos frecuentes, lo que hace que reconozcamos los compuestos fónicos prototípicos de una

lengua. Así, para el español los fonemas consonánticos serán todos aquellos incapaces de formar, sin una vocal palabras o sílabas (Alarcos, 1986).

La frecuencia con la que los sonidos se repiten caracteriza fonéticamente a una lengua. En español de acuerdo con Alarcos (1986) la frecuencia de fonemas es de 47.30% para las vocales y un 52.70% para las consonantes. El porcentaje de las vocales se distribuye de la siguiente manera: el fonema /A/ 13.70%, /E/ 12.60%, /O/ 10.30%, /I/ 8.60%, /U/ 2.10%. La cifra para los fonemas consonánticos se reparte entre fonemas líquidos: 12.80% y 39.90% para los demás fonemas consonánticos. El fonema más frecuente entre los consonánticos fue la /S/ ocupando el 8.00%.[15]

La sílaba es definida como una unidad natural de todas las lenguas, que tiene como núcleo una vocal y como margen una consonante. Las vocales, al ser nucleares, pueden en ocasiones considerarse como sílabas completas (Lara, 2006). Observamos un ejemplo de este fenómeno con la palabra “arroz” donde su separación silábica es “a-rroz”.

La parte de la silaba en la que se realiza el acento es el núcleo o parte central que en español es siempre un fonema vocálico y el resto de la sílaba está constituido por fonemas avocálicos conocida como parte marginal, en donde sólo las vocales /i/, /u/ son posibles (Alarcos, 1986). Por ejemplo una palabra como “cuerda”, puede fragmentarse en dos sílabas “cuer”- “da” la sílaba “cuer”, tiene como núcleo silábico a la vocal “e”, la parte marginal sería “cu”, “r”. Este mismo fenómeno se presenta en todas las palabras diptongadas del español en donde las combinaciones de vocales, dentro

[15] Para obtener información detallada sobre el resto de los fonemas consonánticos consultar (Alarcos, 1986 p.197-200).

del diptongo, siempre serán una vocal abierta (que fungirá como núcleo) y una cerrada (i ó u). Por ejemplo, en la palabra “peine” la sílaba “pei” tiene como núcleo a la vocal “e” y como parte marginal “p”, “i”.

De acuerdo con Lara (2006) el conjunto limitado de formaciones silábicas en una lengua es denominado patrón canónico de la sílaba. Este patrón determina que algunos fonemas solo pueden ocupar ciertos lugares dentro de la sílaba; la sucesión de sílabas, en ocasiones, varía por la posición de la sílaba acentuada dentro de una palabra: inicial, media o final. De acuerdo con Lara (2006) en español las sílabas más frecuentes de manera cuantitativa del patrón canónico son las que presentamos a continuación. Cabe señalar que en el cuadro siguiente se diferencian las sílabas, tanto por su estructura (presencia de consonantes y vocales) así como por la posibilidad de presentar acentuación prosódica. Esta última se marca en el cuadro mediante los apóstrofes que anteceden a la vocal acentuable:

ESTRUCTURA	EJEMPLO
C`V	co.*lé*.gio
CV	*ca*.be.*za*
C`VC	*cóm*.pu.to
V`C	*ác*.to
V`	á.la
CVC	pa.rá.*dos*
V	*u*.sár

Nota: C representa consonantes; V, cualquier vocal, la separación silábica es representada con puntos, la acentuación es prosódica, la estructura silábica ejemplificada está en letra itálica (Lara, 2006).

Dado que el análisis de una lengua se realiza a partir del uso que hacen de ella los hablantes, Serra, Serrat, Solé Bel y Aparici (2000) consideran que existen diferencias en la frecuencia de parición silábica del español si se diferencian hablantes

adultos de niños. La composición de sílabas más frecuentes, presentada en orden decreciente, considerando a los hablantes por su edad, es la siguiente:

Adultos	Niños
CV	CV
CVC	V
V	CVC
CCV	VC
VC	CVV
CVV	CCV
CCVC	

Características semánticas de las palabras

Además de conocer los aspectos fonológicos de la *palabra,* Lara (2006) propone que definir el significado de una *palabra* es parte central de la naturaleza del término. El reconocimiento de *la palabra* como unidad de denominación permite realizar un análisis lingüístico de la misma. Así por ejemplo, al narrar un suceso, los hablantes de una lengua son capaces de dar cuenta del significado de las palabras que utilizan o de reflexionar sobre ellas para aclarar alguna duda proveniente del receptor como "¿a qué te refieres con X?"

La forma representante de la palabra que utiliza la lexicografía como unidad de análisis se denomina *vocablo*. El término *vocablo* es utilizado socialmente como sinónimo de palabra, pero en lingüística su significado difiere. Dentro de la lexicografía el término *vocablo* se utiliza como unidad de análisis de una lengua, su función es representar un símbolo y su uso permite objetivar a la palabra como tema de reflexión

(Lara, 2006). De acuerdo con Lara (2006), la lingüística descriptiva caracteriza a los vocablos de la siguiente manera:

a) Tienen un número entero de sílabas que se ajustan a la función demarcativa de los fonemas. Es decir en un vocablo como “abuelo” encontramos tres sílabas completas (a eso se refiere el número entero de sílabas) aunque difieren en estructura a-bue-lo, la función demarcativa de las sílabas en español es posible gracias a las vocales que hacen el corte al interior de la palabra; En el ejemplo anterior “a”, “e” y “o” marcan los límites silábicos.

b) Tienen el menor número de morfemas flexionales o derivativos que le permitan poseer un significado gramatical pero sin que añadan un significado específico. En el español, las sílabas finales de los vocablos por lo general nos permiten observar este fenómeno. Siguiendo con el ejemplo anterior en “abuelo” la última sílaba “lo” encierra en “o” información morfológica sobre el género y número de este sustantivo.

c) Representan todo el paradigma de flexiones, derivaciones o conjugaciones que se forman a partir de una raíz o núcleo morfemático. En el ejemplo de “abuelo” la raíz o núcleo morfemático “o” podría intercambiarse por “a” (volviendo al vocablo femenino) por “os” (plural) o “as” (femenino y plural); “ito”(diminutivo), etc.

Los vocablos derivados requieren para su estudio de una unificación que permita agruparlos. Como lo mencionamos en la página 54, esta es la función de lematización. Para el español los sustantivos se lematizan en la derivación masculina, singular. Por ejemplo, “gatos”, “gata”, “gatitos” estarían bajo la cabeza lemática “gato”. En el caso de los verbos, la lematización de las flexiones verbales recae sobre su forma infinitiva del

vocablo por ejemplo, “corrían”, “corrió”, “corrimos”; quedarían lematizados bajo el vocablo “correr”

Una vez que se establecen los vocablos de una lengua, en su forma lemática, puede identificarse cuál es el léxico de la misma. Fuera del campo de la lingüística, el léxico de una lengua, es producto de una memoria individual y colectiva construida a partir de la experiencia que tenemos con el mundo. La recolección de datos léxicos es un propósito de la lexicografía, es una labor compleja debido a la cantidad ilimitada de vocablos que un sujeto puede poseer (Lara, 2006). Aunque no se puede determinar el léxico total de una persona o comunidad lingüística, sí se puede dar cuenta, de la cantidad de vocablos que un sujeto necesitaría para poder hablar una lengua y comunicarse en una situación de necesidad. El constante mínimo de léxico con que se puede hablar una lengua es conocido como “vocabulario fundamental” (Lara, 2006).

Cabe mencionar que dependiendo de la base del corpus que se tenga, se puede identificar el vocabulario fundamental no solo de una lengua total, sino incluso de sus variantes dialectales. Así, se podría identificar por ejemplo, el vocabulario fundamental de la variante de los juristas o de los médicos.

En la presente tesis buscaremos identificar el vocabulario fundamental de los sustantivos de la variante escrita presente en textos dirigidos a niños en sus primeros años de escolaridad básica.

Las Palabras Escritas en los Estudios Psicolingüísticos

Dado que la presente tesis se realiza dentro del área de la psicolingüística, la justificación del estudio que en ella realizamos requiere de que presentemos la

importancia que el estudio de las palabras escritas ha tenido en dicho campo: la conciencia que los niños tienen sobre estas unidades lingüísticas y la trascendencia que éstas pueden tener sobre el proceso de alfabetización.

Comenzaremos señalando que la lingüística moderna, durante mucho tiempo, concibió a la escritura como deformación de la oralidad y asumió que no era más que una transcripción de la forma de expresión de una lengua (Lara, 2006). Al describir el papel que la lingüística ha otorgado a la escritura a lo largo de la historia, Zamudio (2010:16) menciona:

> *A pesar de que existen diferentes formas de escribir el lenguaje, la escritura se piensa como única y es siempre la copia. No importa cuáles sean sus unidades de representación: logogramas, silabogramas o letras, ésta sólo traspone al medio gráfico los elementos o unidades que componen el lenguaje oral.*

El español, así como muchas lenguas europeas y algunas amerindias, pertenece al sistema de escritura fonográfico porque simboliza unidades de segunda representación. De acuerdo con Lara (2006:115) está constituido por los siguientes elementos:

> a) un conjunto de trazos visibles en una superficie de contraste, socialmente instituido
>
> b) un conjunto de reglas de representación de formas lingüísticas
>
> c) un conjunto de reglas de complexión del propio sistema
>
> Los trazos del sistema de escritura latino, primer elemento de nuestro sistema, fueron heredados por todas las lenguas del Occidente europeo y se han ido

modificando a través de la historia. Los trazos que constituyen el sistema de escritura se denominan letras y pueden trazarse de distintas formas sin cambiar su significado. Por ejemplo: A, a, a. En cambio, se denomina grafía a cada letra o conjunto de letras que corresponden a un fonema o grupo de fonemas. Por ejemplo el fonema /K/ del español se representa en nuestro sistema de escritura por medio de las grafías <C><K> y <QU> respectivamente.

El segundo elemento refiere a dos reglas que tiene nuestro sistema. La primera, es que se rige principalmente por un principio fonológico que consiste en que a cada fonema le corresponda una grafía en relación biunívoca. Cabe señalar que en español esta regla en términos generales se cumple. Sin embargo, debido a su ortografía, es posible también que para un mismo fonema se empleen diferentes grafías. Por ejemplo /s/ puede graficarse como “z” (en zanahoria), “s” (en salsa) o “c” (en cielo). Así mismo, una grafía puede ser empleada para representar diferentes fonemas. Por ejemplo, “g” en “gorila” no representa lo mismo que en “gis”. Otra excepción del español lo constituyen las grafías que no refieren a fonemas, por ejemplo “h” (en hijo) o “u” en (queso). Finalmente encontramos también dígrafos que refieren a un solo fonema. Por ejemplo, “ch” (en chicle). Y viceversa, grafías que remiten a dos fonemas; por ejemplo “x” (en examen).

La segunda regla es que el orden de escritura y de lectura es de izquierda a derecha y de arriba a abajo.

El tercer elemento del sistema de escritura fonográfico lo constituye el conjunto de reglas de complexión del sistema, que forman la ortografía misma y que determina su correcta escritura.

A pesar de que la ortografía del español se conformó bajo una mezcla de criterios es considerada sencilla y regular por lo que se le contempla dentro del grupo de escritura con ortografía transparente, en comparación con otros idiomas como el francés y el inglés que tienen una ortografía opaca o profunda, regulada principalmente por representaciones que se alejan de los fenómenos fónicos.

Zamudio (2010) muestra que en la ortografía inglesa y francesa las combinaciones ortográficas son indicadores de pronunciación de las palabras. En estas lenguas la mayoría de las palabras que carecen de pronunciación tienen una importante referencia morfológica en el medio escrito como indicar el género de nombres y adjetivos o la persona en el verbo, lo que ayuda a reafirmar la idea de que la relación fonema-letra en estos idiomas no es predominante.

En lenguas como el español la escritura es determinante para reflexionar y analizar la *palabra,* ya que da permanencia a la lengua oral que de otra forma pasaría desapercibida. Solo en las últimas décadas, algunos lingüistas y psicolingüistas han empezado a interesarse en la escritura como un fenómeno complejo (Lara, 2006). Ya no sólo están interesados en especificar los procesos de adquisición de la lengua oral. Dada la importancia de la escritura en nuestra cultura, la alfabetización ha tomado un foco importante para esta área de estudio. Es en este foco que nuestro estudio cobra relevancia.

La alfabetización inicial puede abordarse desde diferentes perspectivas teóricas. Si asumiéramos que leer antecede al aprendizaje de la escritura (Treiman, Bryant, Goswami, 1990), preguntarse por cuáles son las palabras que primero se pueden leer o identificar resulta una pregunta importante. Existen estudios que analizan los procesos cognitivos que las palabras generan: reconocimiento visual y auditivo de palabras, descomposición de segmentos gráficos y sonoros, adquisición de la noción palabra, etc. Cabe señalar que la mayoría delos trabajos sobre lectura o identificación de palabras, se han realizado considerando al inglés del que se cuentan con listados de frecuencia de aparición de las palabras escolares desde el Siglo XIX. Sin embargo, no existe una tradición equivalente en nuestra lengua, son muy escasos los estudios sobre aprendizaje de la lectura y nulos los listados de palabras escritas frecuentes al alcance de los nuños.

Dentro de este tipo de estudios, contar con un corpus de vocablos frecuentes ha comenzado a ser importante debido a que actualmente se analiza la relevancia que tienen las palabras escritas y los elementos que las conforman: categoría gramatical, extensión, frecuencia, letras, sílabas y estructura silábica.

Estudios relacionados con las características gramaticales de las palabras escritas

Una de las características más importantes de una palabra, es la categoría gramatical a la que pertenece. Gracias a los estudios en psicolingüística conocemos que las palabras de una lengua no constituyen listas arbitrarias y sin relación entre sí. En el lexicón o diccionario mental donde se almacenan las palabras, las palabras se agrupan en clases que comparten rasgos morfológicos, rasgos sintácticos y rasgos

semánticos (Giammatteo y Albano, 2009). Adicionalmente, se pueden distinguir dos clases de palabras: léxicas (o de contenido) y de función. Como palabras léxicas se consideran al sustantivo, adjetivo, verbo, adverbio y preposición. Como palabras de función hallamos a los determinantes y a las conjunciones.

Los estudios en psicolingüística muestran que el desarrollo de la conceptualización de *la palabra* está relacionado con la categoría gramatical a la que las palabras escritas pertenecen. Este proceso es estudiado por Ferreiro y Gómez Palacio (1982) a través de una serie de entrevistas tipo clínico a niños no alfabéticos. Las investigadoras descubrieron que en un primer momento los niños reconocen que solo los sustantivos se representan. Durante este periodo parecen asumir que solo el contenido referencial del mensaje está dentro del texto pero no su forma verbal.

Posteriormente, los niños consideran que los sustantivos y el verbo, que ya pueden aislar de una emisión oral, están representados en los segmentos escritos. Los términos funcionales no se consideran como palabras y se rechaza que las cadenas gráficas menores de tres letras puedan decir algo. Por último, observaron que los niños llegan a una etapa más evolucionada donde asignan significado a todas las cadenas gráficas incluyendo las palabras funcionales.

Además de la influencia de la categoría gramatical en la conceptualización de *las palabras,* se ha estudiado que el reconocimiento de una *palabra* escrita varía dependiendo de la categoría a la que pertenece. De acuerdo con Gombert (1990 en Cuetos 2010), dentro del reconocimiento de palabras existe mayor facilidad para

distinguir los sustantivos, verbos y adjetivos y una gran dificultad para aceptar que artículos, conjunciones, preposiciones y otros elementos de enlace sean palabras.

Por su parte, los estudios de Rayner (1975); Rayner y McCinkie (1976 en Cuetos, 2010) señalan a través de estudiar los movimientos oculares en la lectura de textos, que los puntos de destino a los que se dirigen los ojos son las palabras de contenido (palabras plenas) más que a las de función.

Por lo anterior, el análisis del corpus de esta investigación se centra en los sustantivos por ser palabras léxicas (contenido pleno) considerando que son las primeras palabras que los niños en proceso de alfabetización conceptualizan y a las que mayor importancia se les otorga como partes escritas dentro de un texto.

Estudios relacionados con la frecuencia de aparición de las palabras escritas

Otra característica importante de una palabra es la frecuencia (número de ocurrencias) con la que aparece en los textos escritos. Trabajos como los de Cuetos, (2010) muestran que la frecuencia de aparición de una palabra dentro del texto escrito influye en su reconocimiento. Este autor propone que todas las palabras tienen un umbral de activación, que disminuye cada vez que se reconocen dentro de un texto. Considera que los niños comienzan a leer las palabras utilizando una regla de conversión grafema-fonema, pero que a medida que van leyendo una y otra vez la misma palabra, van formando representaciones mentales que ayudan a leerlas sin necesidad de aplicar esta regla. Así, si un sujeto trata de leer una palabra de baja

frecuencia de uso, que no forma parte de su vocabulario visual, la dificultad que presente será mayor que si se enfrentara a una palabra conocida.

Así mismo, la frecuencia de una palabra es considerada una característica importante en los estudios de Defior, Justicia y Martos (1996, en Zamudio 2010) quienes, al comparar el desempeño lector de niños que leen español con el de lectores ingleses, descubrieron que la frecuencia de una palabra modifica la eficiencia en lectura independientemente del idioma en el que lean. Los lectores de su estudio utilizaron la ruta léxica (lectura de palabras completas) cuando se encontraron ante palabras frecuentes y la ruta fonética (subléxica, leer fonema por fonema) cuando se enfrentaron a palabras largas, poco frecuentes, pseudo-palabras, o palabras que no correspondían con la estructura de su idioma.

Estudios relacionados con la complejidad silábica de las palabras escritas

Los estudios de alfabetización inicial han puesto en evidencia que la sílaba tiene prioridad como unidad lingüística en los análisis que los niños son capaces de realizar. Desde una perspectiva cognoscitivista realizando estudios sobre la alfabetización de niños angloparlantes, Treiman, Bryant y Goswami (1990), señalan que la identificación inicial de las palabras está relacionada con la estructura silábica y/o morfémica de las mismas**.** La dificultad para reconocer una palabra escrita radicaría en que presentara combinaciones de letras menos usuales y/o menos transparentes para un lector. Desde una perspectiva similar, Ehri (1975) reportó que ante la consigna que daba a los niños de dividir enunciados en palabras, los pre-lectores distinguían más bien sílabas o agrupaban palabras de acuerdo con los patrones tonales. Asimismo, los estudios de Liberman, Shankweiler, Fischer y Carter (1974, en Zamudio, 2010) mostraron cómo

niños preescolares y de primer año, en contextos meramente orales (desprovistos de escritura), reconocieron sílabas con mayor facilidad que fonemas.

Desde otra perspectiva teórica, para Ferreiro (2002) la sílaba tiene para los niños una identidad psicológica indiscutible por lo menos desde los 4 o 5 años. En estudios sobre la representación alfabética, Ferreiro y Gómez Palacio (1982) constatan que, en etapas avanzadas del proceso, la sílaba cobra importancia en el análisis alfabético. Parece haber una jerarquía en el logro de representación de la sílaba. La representación comienza con la estructura consonante-vocal (CV), luego con la estructura consonante-vocal-consonante (CVC) y posteriormente con la estructura (CCV).

De acuerdo con Quinteros (1997), los niños pre-alfabéticos (entre los niveles silábicos al alfabético) se centran fuertemente en la sílaba en sus intentos de representación escrita. Desde esta perspectiva, Cano y Vernon (2008) indagaron el uso que los niños en proceso de alfabetización hacen de las consonantes. Se percataron de que en general los sujetos denominaban las letras usando el segmento fónico que las representaba. Pero también fue muy frecuente el uso del nombre de la letra o la sílaba que iban a escribir o a identificar. También en los estudios de Calderón (2010) se reportó que los niños pueden acceder a la sílaba sin previo conocimiento de la escritura. Detrás de estos estudios se entiende que en las lenguas romances, como el español, la fonetización de la escritura es silábica. Por lo tanto, cuando los niños empiezan a anotar las sílabas conformadas por una consonante y una vocal no significa que puedan escribir todas las clases silábicas de manera convencional.

Estudios relacionados con la longitud gráfica las palabras escritas

Otra característica propia de la palabra escrita es su longitud. Existen investigaciones como la de Rayner (1977) y Just y Carpenter (1980 en Cuetos, 2010) que reportan que las palabras largas producen mayores pausas en la lectura que las cortas. De acuerdo con Cuetos (2010), para el español la lectura de palabras largas (tres o más sílabas) podría ser más difícil que la de palabras cortas (una o dos sílabas). En inglés, la lectura de palabras con más de cinco caracteres, independientemente del número de sílabas involucradas, presentaría mayores dificultades.

Estudios relacionados con la similitud visual de las palabras escritas

Los trabajos de Cuetos (2010) reportan que los seres humanos poseemos un área en el cerebro (detectable mediante neuroimagen) en el que almacenamos a las palabras de manera visual; en la medida que aumenta la capacidad lectora de un sujeto se modifica la organización cerebral permitiéndole reconocer con mayor facilidad las palabras. De acuerdo con este autor, un requisito para poder utilizar la vía léxica en la lectura de una palabra, es disponer de su representación ortográfica porque, de lo contrario, cuando se trate de una palabra de baja frecuencia de uso, que no forme parte del vocabulario visual del lector, éste no las podrá leer. Goswami (2006 en Cuetos 2010), reconoce que esta zona se encuentra en el área 37 del cerebro denominada área visual de la forma de las palabras.

Estos autores consideran que la destreza lectora varía en cada sujeto dependiendo del corpus de palabras que puede identificar visualmente con rapidez. De acuerdo con este paradigma, el análisis de las letras que conforman una palabra solo es necesario para la identificación de palabras poco frecuentes.

Por su parte, y desde una perspectiva teórica diferente, Cano y Vernon (2008) al analizar los errores cometidos por los niños al escribir palabras reportaron que los niños cometen equivocaciones inteligentes basadas en la semejanza gráfica de las letras. Por ejemplo, al escribir palabras que inician con la letra “B” como “BUZO”,”VENADO” (atendiendo al criterio de sonido que representan, aunque por ortografía se escriban diferente) un porcentaje alto de niños eligió a las consonantes “P”,”D” y “F” en lugar de “B”

Desde una postura también psicogenética, los estudios de Ferreiro (1997), Quinteros (1997) y Ferreiro y Zamudio (2008) se han interesado por estudiar las decisiones gráficas que los niños prealfabéticos toman al momento de escribir "palabras". Estas autoras no han atendido a la familiaridad que tienen los niños con las formas escritas de las palabras (con palabras específicas más frecuentes) porque el centro de su análisis ha sido la lógica que los niños imprimen al funcionamiento del sistema de escritura. Si bien no es propósito de esta tesis realizar estudios psicolingüísticos específicos, nos preguntamos si esta variable pudiera, en algún sentido, condicionar el desempeño escrito de los niños. En este sentido notamos la posibilidad de incluir la variable frecuencia de palabra en la selección de las palabras que conforman los instrumentos empleados, así como algunos rasgos gráficos y fonológicos de éstas: longitud de las palabras y constitución silábica. Como lo desarrollaremos con profundidad en el próximo capítulo, el propósito de esta tesis es ofrecer un recurso metodológico útil para la realización de trabajos psicolingüísticos en español que tengan como foco de análisis la actividad cognitiva que pueda realizarse sobre la palabra gráfica.

CAPÍTULO 3

METODOLOGÍA

Como se pudo observar en los capítulos anteriores, el estudio con corpora ha facilitado el tratamiento de información y la profundidad con la que se pueden realizar análisis lingüísticos. A través de la presente investigación, pretendemos aportar un instrumento metodológico que posibilite continuar los estudios psicolingüísticos, sobre todo en lo referente a alfabetización inicial, considerando como variable de análisis a las palabras escritas. Para ello nos hemos planteado como propósito de la presente tesis la construcción de un corpus especializado que dé cuenta de cuáles son los sustantivos más frecuentes a los que los niños mexicanos en momentos iniciales de la alfabetización están expuestos de manera visual. En consecuencia, nos propusimos responder a las siguientes preguntas:

¿Cuáles son los sustantivos escritos más frecuentes en textos para niños mexicanos en momentos iniciales de la alfabetización? y ¿qué estructuras silábicas y qué contextos resultan más frecuentes entre las palabras que integran el vocabulario fundamental (sustantivos escritos)?

Como lo mostraremos más adelante, para la realización del presente estudio consideramos textos escritos de amplia difusión entre la población infantil. Si bien en la presente tesis nos dedicaremos exclusivamente a los sustantivos comunes, tenemos como una segunda intención de nuestro trabajo dejar lista una base de datos que pueda analizarse con mayor profundidad en estudios posteriores.

El estudio que realizamos se caracteriza por ser exploratorio y descriptivo. Se considera exploratorio porque buscó construir una herramienta para continuar con las investigaciones sobre la adquisición de la lectura y la escritura en contextos hispanoparlantes. En este sentido, como menciona Dankhe (1986, en Samaja, 2007) esta investigación no constituyó un fin en sí misma sino que relacionó variables y establece el "tono" para investigaciones posteriores. También se considera un estudio descriptivo, porque midió y describió con la mayor precisión posible dimensiones de un fenómeno (palabras escritas). Para ello, caracterizamos un corpus general de palabras presentes en el ámbito infantil y analizamos, a partir de él, las características de los sustantivos más frecuentes utilizando la estadística descriptiva como herramienta de análisis.

De lo anterior se derivó la necesidad de conformar primero un corpus general de palabras escritas para el público infantil para después realizar un corpus específico de sustantivos frecuentes. A continuación exponemos las pautas que guiaron la elaboración de nuestro corpus.

Construcción del Corpus General

La construcción de un corpus requiere ser regulada por una metodología específica de manera que constituya un instrumento válido que represente a la lengua de una comunidad. Como lo dimos a conocer en el Capítulo 2, la lexicografía de corpus ha establecido criterios para estudiar el léxico de manera científica de forma que el resultado de sus estudios sea válido y generalizable.

A lo largo de nuestra investigación hemos indagado e incorporado los avances de esta metodología a nuestro estudio. Un parámetro fundamental que propone la lexicografía de corpus es la labor de documentación (Sinclair, 2005). Entendiendo a este principio como la necesidad de determinar la estructura de un corpus con claridad, hemos tratado con especial cuidado de cumplirlo. En este capítulo detallaremos todo el procedimiento de elaboración de nuestro instrumento. Mostraremos los criterios de selección de la información que conformó el corpus y el proceso de tratamiento de la información. Así mismo, daremos a conocer los argumentos que justifican las decisiones que tomamos sobre su elaboración.

Los textos que conforman la base de datos

Los textos que sirvieron como fuente de información, fueron los libros de lectura oficiales para niños de primer y segundo año de primaria compilados por la Secretaría de Educación Pública a través del programa nacional para el fortalecimiento de la lectura y la escritura en la educación básica, editados para los ciclos escolares 1997-1998 y 2012- 2013.

El conjunto de libros forma parte de los libros de texto nacionales. Se trata de documentos elaborados por el gobierno de la República y se entregan de forma gratuita a todos los niños que cursan primer y segundo grados en las escuelas primarias del país. Su uso se encuentra reglamentado en los programas educativos de cada grado escolar como material complementario al libro de actividades de la materia de español, por lo que suponemos que la gran mayoría de niños en México tiene acceso a estos

materiales dentro de su vida escolar diaria. Seleccionamos este material tomando en cuenta dos criterios: la cantidad de tiraje editorial y los títulos contenidos en cada libro.

Respecto al tiraje editorial, los libros de la Secretaría de Educación Pública en México cubre 3’398,000 ejemplares por edición, lo que representa el mayor tiraje editorial de materiales escritos en el mundo (SEP, 1997). Esta cantidad de impresiones hace que los libros de lectura oficiales sean considerados como tomos representativos de la literatura infantil a la que los niños están expuestos en su vida cotidiana.

Seleccionamos seis libros de lectura, tres correspondientes a primer grado y tres a segundo. Compilados, forman un listado de 604 títulos. Están compuestos por tres géneros correspondientes al uso culto del español mexicano: literatura, textos informativos y textos epistolares. La Tabla 1, muestra la distribución del total de lecturas de acuerdo con su género textual.

Tabla 1.
Porcentaje de títulos por género textual

Géneros textuales	Número de títulos	Porcentajes
Informativos	76	12.60
Literarios	526	87.00
Otros	2	0.33
Total	**604**	**100.00**

En la información anterior podemos observar que los textos literarios predominan dentro de nuestra base de datos. De acuerdo con Lara, Ham y García (1979), la literatura representa a la lengua culta como nivel de uso del español mexicano. Este nivel elevado del español estándar posee dos características importantes. La primera es que contiene un vocabulario intelectualizado y rico. La segunda es que contiene una

sintaxis rica que, por lo tanto, sirve como modelo de corrección. Actualmente, sabemos que específicamente la literatura infantil busca potenciar en los niños el aprendizaje acerca del mundo y la cultura en la que se desarrollan. Algunos de los títulos contenidos en los libros de lectura de la SEP son de carácter universal, mientras que otros representan parte de la tradición nacional. El siguiente cuadro muestra algunos títulos presentes en nuestra base de datos a modo de ejemplo.

Títulos de literatura universal	**Títulos de literatura nacional**
El convite del zorro y la cigüeña	El señor don Gato
El traje del emperador	El paseo Chapultepec
El gato con botas	Lola Álvarez Bravo: cazadora de imágenes
Romeo y Julieta	Corrido de Pancho Villa
Ricitos de oro y los tres osos	Fundación de México Tenochtitlán
Teseo y el Minotauro	Refranero popular
La hormiga y la cigarra	

Los títulos enlistados en el cuadro anterior, además de estar presentes dentro del currículum nacional, se producen y editan con regularidad en casas editoriales privadas. Por ello consideramos que simbolizan una parte significativa de la tradición cultural escrita de nuestro país.

Los textos que conforman la base de datos son la materia prima para generar un corpus. Sin embargo, la calidad del mismo no depende únicamente de la selección de textos escritos, sino del propósito que justifica su creación y el cuidado que se tiene durante el proceso de elaboración. Las decisiones que guiaron la elaboración de nuestro corpus corresponden con las de un corpus especializado que parte de considerar textos completos (su foco son los sustantivos escritos presentes en textos dirigidos a niños de 1º y 2º años de primaria). Es también un corpus cerrado (porque

una vez finalizado no admite la inclusión de más palabras) y monolingüe (porque sólo se incluyeron versiones en español de los libros de texto involucrados.

Herramientas para la elaboración del corpus general

La creación del corpus de sustantivos más frecuentes en textos escritos para niños mexicanos atravesó diferentes fases. En un primer momento, creamos la base de datos y elegimos el analizador léxico más pertinente de acuerdo con los objetivos del estudio.

Después, una vez seleccionados los textos que conformarían la base de datos, creamos una base de datos digital. Para ello, fue necesario buscar en línea los textos contenidos en cada libro de lectura y verificar que los textos coincidieran con el texto original. En algunos casos, las lecturas digitales tenían modificaciones, sobre todo en el uso de nombres propios, por lo que fue necesario comparar el texto digital con el original y realizar modificaciones de forma manual. Con este procedimiento, corroboramos que el contenido de los textos en ambas fuentes (texto impreso y digital) fuera el mismo. En caso de que las lecturas no estuvieran disponibles en formato digital las transcribimos, cuidando conservar el texto en formato original.

De manera simultánea a la elaboración de la base de datos, probamos diferentes analizadores léxicos[16]. En el mercado de software, existen diferentes analizadores léxicos; de manera que tuvimos que realizar la comparación de cuatro de ellos para tomar la decisión de cuál emplear para los fines de esta tesis. Los cuatro

[16] Para conocer la descripción de estas herramientas, remítase al capítulo 2.

programas que comparamos fueron: a) Microconcord[17]; b) Antconc[18]; c) UAM Corpus Tool [19]; y d) Word Smith Tools.

Seleccionamos el analizador de la Universidad de Oxford, Word Smith Tools versión 6.0, en adelante (WST), debido a que posee todas las herramientas básicas para realizar estudios lingüísticos[20]. El programa WST, posee una versión gratuita en línea que nos ofreció los beneficios de un estudio piloto ya que realiza funciones similares a la versión comercial a excepción de la capacidad de palabras que puede analizar. La versión gratuita sirvió para correr los datos iniciales y verificar que fungieran como un apoyo para lograr los objetivos de la investigación. Zapata (2005) refiere que un instrumento es válido en la medida que logra medir lo que el estudio pretende. En este sentido, WST resultó ser un instrumento pertinente para esta investigación por las herramientas que posee.

---

[17] Elaborado por Oxford University Press que utiliza, como base de datos, el Birmingham Collection of English Texts. Incluye 40'000,000 palabras (mismas que sivieron como base de datos del Cobuild Dictionary dirigido por J. Sinclair). Este programa tiene la opción de encontrar y analizar concordancias de palabras y analizar el contexto en el que aparecen.

Las dos colecciones de textos que ofrece pertenecen a dos registros específicos del inglés. La primera (Corpus Collection A) está dedicada al inglés periodístico y consta de cinco corpora de 200,000 palabras, cada uno de ellos referente a una sección distinta de la organización de un periódico (noticias nacionales, noticias internacionales, economía, cultura, y deportes). La segunda (Corpus Collection B) consta de cincocorpora de inglés académico, distribuidos en las siguientes áreas: físicas y biología, medicina y psicología, historia y derecho, filosofía y letras, y religión.

[18] Software desarrollado por el profesor Laurence Anthony del Center for English Language Education in Science and Engineering (CELESE), School of Science and Engineering, Waseda University en Japón. El programa, ofrece concordancias, listas de palabras más frecuentes, colocaciones, agrupaciones, n-gramas y lista de palabras clave.

[19] Elaborado por Michael O'Donell (2008) es un conjunto de herramientas para la anotación lingüística de los textos. Permite construir un corpus junto con los esquemas de anotación para los archivos de texto. Cada análisis debe ir por niveles. Encontramos dos tipos de anotación: 1) anotación de documento (considera la totalidad del texto) por ejemplo, género textual, características del escritor y 2) anotación de segmentos (cláusulas, sintagmas). Utilizando una herramienta gráfica el usuario puede definir etiquetas apropiadas para cada nivel de anotación. También posee un etiquetador automático basado en la correspondencia de patrones léxicos.

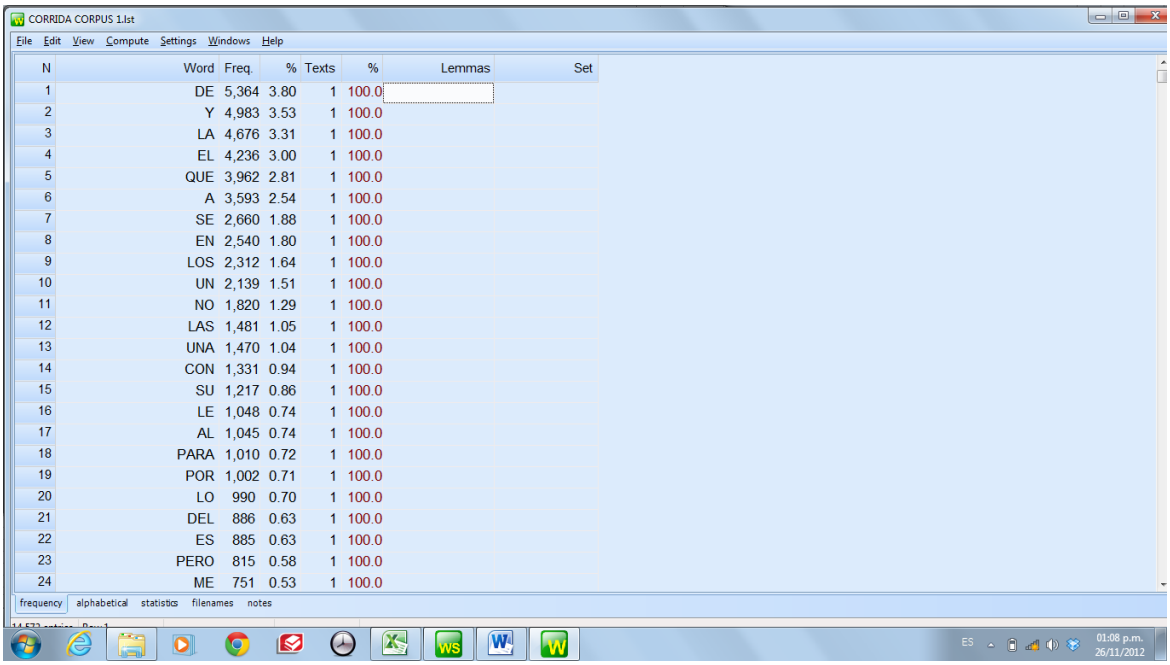
[20] Ver el apartado de analizadores léxicos en el capítulo 2.

Tomamos la decisión de trabajar con el Word Smith Tools (WST) ya que, además de estar avalado por diferentes lexicógrafos y haber sido utilizado para la creación del diccionario de la universidad de Oxford, permite procesar el texto antes de su análisis y agregar etiquetas morfosintácticas. Así mismo, este programa es muy versátil en la forma de presentar los listados de vocablos para realizar diferentes tipos de análisis. Exponemos a continuación los criterios que nos llevaron a elegir WST: i) Calidad y finalidad de creación. La elaboración de WST fue elaborado por la Universidad de Oxford con el propósito de apoyar la creación de diccionarios, lo que asegura que posee las herramientas básicas para elaborar un corpus; ii) presenta la posibilidad de incluir los textos que el investigador prefiera. A diferencia de otros software, la base de datos puede crearse desde el principio; iii) los datos pueden analizarse en forma global y de manera específica. Si bien WST puede dar cuenta de una base de datos completa, tiene una función "Stop List" que permite crear listas de exclusión. En ella se especifican los elementos que no se quieren estudiar para que los excluya de los resultados que presenta. iv) permite analizar palabras en contexto. Muchos de los software que revisamos tienen la posibilidad de analizar a las palabras en su contexto, pero en su mayoría, son palabras que están cargadas en una base de datos previa e inamovible; v) tiene la posibilidad de trabajar con varias lenguas, entre ellas, el español, a diferencia de la gran mayoría de software que están pensadas para utilizar y analizar solo el inglés; vi) ofrece acompañamiento en línea, tutorial y personalizado, además de proporcionar manuales de uso completos y la posibilidad de contactarnos directamente con el desarrollador, Mike Scott, en caso de requerir ayuda.

Descripción General del Análisis

Para comenzar el análisis general de datos, el programa WST especifica que las listas de palabras iniciales deben estar en formato .txt, por lo que exportamos los textos contenidos en los libros de lectura al formato determinado.

Con los textos en formato.txt y la herramienta *Word List* creamos la base de datos para el corpus general (totalidad de palabras, frecuencia, concordancias, etc.). Al realizar la primer corrida de datos, *WST* detectó como error la presencia de números y guiones, por lo que fue necesario eliminar del archivo .txt estos elementos. Una vez que la base de datos se depuró corrimos nuevamente el analizador, obteniendo así el primer análisis de palabras completo. La Figura 1 muestra la pantalla final del primer análisis de datos con las 24 primeras palabras del corpus general ordenadas de acuerdo con su frecuencia.



N	Word	Freq.	%	Texts	%	Lemmas	Set
1	DE	5,364	3.80	1	100.0		
2	Y	4,983	3.53	1	100.0		
3	LA	4,676	3.31	1	100.0		
4	EL	4,236	3.00	1	100.0		
5	QUE	3,962	2.81	1	100.0		
6	A	3,593	2.54	1	100.0		
7	SE	2,660	1.88	1	100.0		
8	EN	2,540	1.80	1	100.0		
9	LOS	2,312	1.64	1	100.0		
10	UN	2,139	1.51	1	100.0		
11	NO	1,820	1.29	1	100.0		
12	LAS	1,481	1.05	1	100.0		
13	UNA	1,470	1.04	1	100.0		
14	CON	1,331	0.94	1	100.0		
15	SU	1,217	0.86	1	100.0		
16	LE	1,048	0.74	1	100.0		
17	AL	1,045	0.74	1	100.0		
18	PARA	1,010	0.72	1	100.0		
19	POR	1,002	0.71	1	100.0		
20	LO	990	0.70	1	100.0		
21	DEL	886	0.63	1	100.0		
22	ES	885	0.63	1	100.0		
23	PERO	815	0.58	1	100.0		
24	ME	751	0.53	1	100.0		

Figura1. Primer Análisis del Corpus General obtenido con la Función Word List del analizador léxico WST

El listado de la primera corrida de datos constó de 14,544 elementos. El objetivo específico de este trabajo fue caracterizar exclusivamente a los sustantivos más frecuentes. Para ello, requerimos excluir las palabras pertenecientes a esta categoría gramatical del resto de las palabras presentes en el corpus general, proceso que describimos a continuación.

Construcción de listas de exclusión y clasificación de palabras por categoría gramatical

Dentro de este trabajo, no solo nos interesó obtener del corpus general la lista de palabras más frecuentes, sino también deseábamos conocer los contextos silábicos que conforman a los sustantivos. Para llevar a cabo este análisis, creamos una lista exclusiva con las palabras pertenecientes a esta categoría gramatical. La exclusión de palabras se realizó con ayuda de dos herramientas, *Stop List* de *WST* y Excel. Con ayuda de las herramientas del programa computacional Excel aislamos todas las palabras sustantivas de las otras categorías.Y con las categorías restantes formamos una lista de exclusión. La lista de exclusión consistió en hacer un listado en formato .txt. con las palabras que se quisieran eliminar del análisis.

La construcción de la lista de exclusión requirió de la clasificación de las palabras de acuerdo con su categoría gramatical. Para asignar cada palabra dentro de una categoría revisamos su contexto de uso en los textos. De acuerdo con Giammatteo y Albano (2009), las categorías posibles fueron dos: palabras léxicas y palabras de función.

Al categorizar a las palabras consideramos la posibilidad que tienen algunas de ellas de pertenecer a dos o más categorías, dependiendo del contexto en el que se utilizan. Para ello, antes de asignar a la palabra dentro de una categoría, verificamos en qué sentido se utiliza dentro del texto original y cómo se categoriza dentro de los diccionarios. Para realizar esta tarea utilizamos dos herramientas. Por un lado, el uso de dos diccionarios, la versión en línea del *Diccionario de la Real Academia Española (REA)* y el *Diccionario del Español Usual en México;* por otro, la revisión de los contextos de aparición de una palabra dentro del texto (concordancias). Para esta tarea, usamos la función *Concord* WST, que nos permitió revisar qué palabras anteceden y suceden a la palabra dentro de una oración.

La herramienta *Concord del WST*, busca una palabra previamente seleccionada dentro de la base de datos general del corpus y muestra en pantalla el contexto oracional de uso de la misma. Por ejemplo, algunos fragmentos textuales que aparecieron en pantalla al seleccionar la palabra “dulce” fueron:

1. Subterráneas que proporcionan agua dulce. ¿Cómo nos aguantamos sobre

2.lo que descubrió.Las tortugas de agua dulce son animales pequeños. Viven

3. ya conoces un poco la historia del dulce en México. El dulce nos produce

4. ¡Ese quiero! Estiro la mano y tomo un dulce con infinito cuidado para

5. dijo: nuestra madre tiene la voz más dulce. Tú eres el lobo. Después de

En las líneas 1, 2 y 5 la palabra “dulce” correspondió a la categoría gramatical de adjetivos al modificar a los sujetos “agua” y “voz”. En cambio, en los fragmentos 3 y 4 la

palabra “dulce” se utilizó como sustantivo. Por lo tanto, incluimos el término “dulce” dentro de dos categorías: adjetivos y sustantivos.

Al realizar el procedimiento de categorización de palabras, encontramos la existencia de pseudopalabras, pertenecientes a expresiones o juegos de palabras como: “AAHHHHHHHHHHHHHHHH”, “ADO/ADI”, “AGHHH”, “AICNACREM”. Excluimos estas expresiones de la base de datos original, al igual que hicimos con los números y signos, por no pertenecer a las palabras convencionales del español.

Realizamos también una lista de exclusión, bajo el mismo procedimiento, para las categorías de verbos y adjetivos que nos permitió verificar la pertinencia de clasificación de los sustantivos. Por el interés de este estudio, trabajamos solamente con la lista de exclusión para sustantivos y dejamos las otras listas de exlusión para un trabajo posterior.

Una vez que obtuvimos la lista de exclusión para sustantivos, la incluimos en el primer análisis general del corpus con la función *StopList*. Esta función nos permitió obtener el listado de sustantivos y continuar con su lematización.

La lista de exclusión para sustantivos estuvo conformada por 9,149 palabras. Su incluión nos permitió obtener el listado de sustantivos presentes en nuestro corpus. Hasta aquí, hemos descrito el procedimiento de elaboración del corpus general y cómo de él obtuvimos el corpus de sustantivos. En lo que sigue, mostraremos el procedimiento que seguimos para poder obtener un análisis más específico de las palabras sustantivas.

Creación del Corpus de Sustantivos más Frecuentes

Proceso de lematización

De la lista completa de sustantivos formada por 4,577 tipos de sustantivos (types) no todos fueron relevantes para este trabajo. Primero creamos un vocabulario fundamental y de él analizamos únicamente los sustantivos más frecuentes, es decir aquellos con ocurrencia mayor o igual a 20.

Para la creación del vocabulario fundamental utilizamos la herramienta *Lemma* de *WST*. Esta herramienta permitió seleccionar una palabra como entrada léxica y dentro de ella contabilizar la ocurrencia (tokens) de todas las palabras que compartían la misma familia léxica. Por ejemplo, marcamos la palabra “abuelo” como vocablo, por ser singular y masculino e incluimos dentro de él a las palabras *abuelos, abuelita, abuelito, abuelas*. La ocurrencia de este vocablo es de 166 por lo que formó parte de nuestro vocabulario fundamental.

Al lematizar las palabras, observamos que algunas carecían de entrada léxica en la lista de palabras original. En caso de que varias palabras de la misma familia léxica sumaran una ocurrencia mayor o igual a 20 creamos una entrada léxica para ellas. Para llevar a cabo este procedimiento fue necesario exportar la lista de lemas a Word y agregar en este formato el vocablo. Una vez que lematizamos a todos los sustantivos seleccionamos los más frecuentes para conformar el vocabulario fundamental.

Análisis silábicos

Una vez conformado el vocabulario fundamental procedimos al análisis de las sílabas que conformaron cada vocablo. Para ello utilizamos el programa estadístico

SPSS, incluimos en él todas las palabras de nuestro vocabulario fundamental, segmentamos cada palabra en sílabas y asignamos un valor a cada una dependiendo de la posición que ocupó dentro de la palabra y de la estructura silábica que la conforma.

Con las listas de sílabas y la estadística descriptiva observamos la frecuencia de aparición de sílabas dentro de la palabra de acuerdo con su posición, es decir, contabilizamos cuántas sílabas ocuparon primera, segunda, tercera o cuarta posición dentro de un vocablo. Esta categorización de sílabas nos permitió excluir las sílabas en primera posición para realizar un análisis más específico sobre ellas.

En esta tesis nos interesó mostrar las características de las sílabas iniciales por la influencia que estas sílabas ha tenido en diversos estudios relacionados con el desarrollo de la conciencia fonológica, el reconocimiento de palabras y el conocimiento del nivel de escritura que tienen los niños. Así, estudios como los de Alvarado (1997), Quinteros (1997), Vernon y Calderón (1999), Vernon (2002), Calderón (2010) controlan a las letras que conforman las sílaba iniciales en tareas de identificación de letra inicial, escritura de sustantivos comunes con una letra inicial previamente seleccionada u omisión del primer fonema de una palabra.

A partir del listado de sílabas en posición inicial, realizamos varios análisis. Observamos primero cuáles fueron las letras iniciales más frecuentes en esta posición, después clasificamos a las sílabas de acuerdo con su estructura silábica. Esta clasificación nos posibilitó contabilizar la frecuencia de aparición de cada estructura. Realizamos también el análisis específico de cada estructura silábica, enlistando qué

sílabas se presentaron dentro de cada una. A través de ordenarlas alfabéticamente, obtuvimos la frecuencia de aparición de cada letra como inicial. A partir de la sumatoria de ocurrencias de las sílabas en cada estructura, nos fue posible también obtener el porcentaje de aparición de cada letra de nuestro alfabeto por estructura.

Una vez que obtuvimos los porcentajes de aparición de las letras en cada estructura silábica, ordenamos los datos con base en deciles, con los cuales obtuvimos sus intervalos de aparición. En cada intervalo, colocamos la frecuencia de cada letra en base al decil. Para calcular la media de cada estructura, utilizamos la fórmula de datos agrupados propuesta por Triola (2009). Realizar este análisis estadístico, nos permitió describir cuáles fueron las letras más frecuentes y en que contextos silábicos se presentaron. La obtención de estos resultados nos permitió realizar un análisis comparativo con los resultados de investigaciones realizadas en torno al léxico del español.[21]

En el siguiente capítulo, mostraremos los resultados obtenidos de cada fase del procedimiento aquí descrito. Respetamos el orden cronológico del proceso. Primero, mostraremos los resultados del análisis del corpus general para después, describir el corpus exclusivo de sustantivos y la valoración del vocabulario fundamental. Finalmente, mostraremos los resultados del análisis de las sílabas iniciales de nuestro vocabulario fundamental.

[21] Si bien hemos mencionado con anterioridad que las características de nuestro trabajo no son idénticas a ningún trabajo realizado en contextos hispanoparlantes, realizar un análisis comparativo con trabajos similares realizados para el español nos permitió verificar la representatividad y validez del mismo.

CAPÍTULO 4

ANÁLISIS Y DISCUSIÓN DE RESULTADOS

Análisis del Corpus General

Como lo mencionamos en la metodología, el corpus de esta tesis estuvo integrado por 604 textos escritos, dirigidos a niños, compilados en 6 libros de lectura. Digitalizamos todos los textos y con ayuda del programa WST sistematizamos una base de datos que nos permitió caracterizar el corpus general.

Para analizar el listado de palabras que arrojó la compilación de los textos utilizamos tres funciones de este analizador léxico: *Word List, Concord* y *Lemma*. Con la función *Word List*, establecimos la frecuencia de aparición de cada palabra e identificamos la longitud de cada una, de acuerdo con el número de letras que la conforman. Esta función también nos permitió ordenar la lista de palabras en orden alfabético lo que nos posibilitó identificar términos con errores ortográficos o de transcripción. La función *Lemma*, nos ayudó a crear el vocabulario fundamental para identificar y marcar los lemas en los que podrían organizarse las palabras. Por último, la función *Concord* nos mostró los diferentes contextos de aparición de una palabra, con lo que verificamos la pertinencia de clasificación de cada término de acuerdo con su función gramatical dentro de los textos. A continuación mostraremos los resultados obtenidos con el apoyo de cada función.

Al analizar con *Word List* la compilación digital de los 604 textos escritos, extrajimos un total de 141,230 palabras a las que denominamos “ocurrencia total de palabras” (tokens). Dentro de las ocurrencias del corpus general encontramos 14,572

tipos diferentes de palabras (types). Para lograr una descripción más detallada organizamos todas las palabras en cuartiles y observamos: 1) la frecuencia de aparición de los términos, 2) la distribución de tipos de palabra de acuerdo con el número de ocurrencias y, 3) los términos con mayor y menor frecuencia del corpus general.

Análisis del Corpus General por Cuartiles

El primer cuartil de palabras se conformó por 10 palabras diferentes, todas ellas no léxicas, con ocurrencias entre 2,139 y 5,364. Este cuartil contiene a las palabras más frecuentes del corpus con el porcentaje más bajo de tipos diferentes de palabras 0.06%. En el segundo cuartil, hallamos la presencia de 71 palabras con ocurrencias entre 153 y 1,920 lo que representa un .48% de la totalidad de tipos de palabra del corpus general. El tercer cuartil, se conformó por el 6.46% de tipos de palabra al contener 942 palabras con ocurrencias entre 15 a 150. Por último, en el cuarto cuartil encontramos la mayor parte de tipos de palabra 92.97%, en este cuartil encontramos 13,549 palabras con ocurrencias entre 1 y 14.

El término más frecuente fue “DE” con 5,364 ocurrencias (frecuencia de aparición), esta palabra también fue reportada como la más frecuente dentro del corpus CUMBRE (Cantos y Sánchez, 2011). El menor número de ocurrencias (una aparición) se presentó en 7,281 vocablos, lo que representó casi el 50% de la totalidad de tipos de palabras. Esta cifra concordó con los resultados de los estudios comparativos de Rojo (2008). Este investigador, al analizar el porcentaje de hápax legomena (palabras con una sola aparición) presentes en diferentes corpora reportó que las palabras con frecuencia igual a uno se mantienen alrededor del 40% de la totalidad del corpus

independientemente de la cantidad de palabras que éste contenga. Si bien más adelante prestaremos atención a la naturaleza de las palabras del corpus (plenas o de función) y a la categoría gramatical que fungen en los textos, para la descripción del corpus general consideramos también la longitud de las palabras (expresada en el número de grafías que se emplean para su representación ortográfica convencional). A continuación detallamos esta información.

Longitud de Palabras de acuerdo con el Número de Letras que las Conforman

El uso de *Word List* nos facilitó el análisis de la longitud de los términos. Al respecto observamos que hubo 23 posibilidades de conformación de palabras. Encontramos palabras formadas por una sola letra (como "A" u "O"), hasta palabras formadas por 24 letras. La media general de longitud de palabra fue de 4 letras. Sólo el 1% de los vocablos presentaron 12 o más letras[22]. Por su longitud, sobresalieron las palabras representadas con 2 letras por conformar el mayor porcentaje del corpus (23%). La distribución de palabras de acuerdo con el número de letras que contienen en nuestro corpus coincidió con la distribución de palabras del corpus CUMBRE. La Tabla 2 muestra el análisis comparativo de ambos corpora por cuartiles.

[22] Aunque existen términos largos en el español como: "terroríficamente" y "resplandecientes", la mayoría de palabras que tienen entre 16 y 26 letras son juegos de palabras que carecen de significado propio como los términos: "recontraotorrinolaringoleo", "supercalifragilística" y "entamabralinguladita".

Tabla 2.

Número y Porcentaje de Palabras del Corpus General y del corpus CUMBRE por Número de Letras.

	Corpus general		Corpus CUMBRE	
Longitud de palabra	**Ocurrencia**	**Porcentaje**	**Ocurrencia**	**Porcentaje**
1 letra	9,017	6.38	816,275	3.75
2 letras	33,289	23.57	3,642,015	16.72
3 letras	21,889	15.50	4,366,524	20.04
4 letras	15,630	11.07	3,637,621	16.70
5 letras	18,156	12.86	2,408,555	11.06
6 letras	14,578	10.32	1,804,230	8.28
7 letras	11,194	7.93	1,655,552	7.60
8 letras	8,061	5.71	1,150,188	5.28
9 letras	4,511	3.19	830,798	3.81
10 letras	2,714	1.92	533,148	2.45
Más de 10 letras	2,191	1.54	940,426	4.32
Total	**141,230**	**100.00**	**2,096,843**	**100.00**

Como lo mostraremos más adelante, diferenciamos las palabras gráficas "léxicas" de las de "función". Resulta pertinente señalar, por lo pronto, que las palabras de función coinciden en ser las de menor longitud (oscilan entre 1 y 3 letras).

De los 14,572 tipos de palabras (types), obtenidos del primer análisis general, excluimos 398 términos por tratarse de "no-palabras" (onomatopeyas, palabras escritas al revés –"amerc", para "crema" –, términos que simulaban ser de otras lenguas o parte de juegos de palabras. Por ejemplo, "patapáfate", "on, onga". En consecuencia, en lo que sigue del análisis del corpus general, trabajamos con una base total de 14,174 tipos de palabras.

Clasificación de Palabras de acuerdo con su Categoría Gramatical

Una vez que obtuvimos el listado de palabras general, procedimos con la clasificación de las palabras en categorías gramaticales de acuerdo con su uso dentro del corpus. Para ello, como lo mencionamos en el capítulo de metodología, fue necesario utilizar varias herramientas: diccionarios escritos y virtuales y la función Concord, para revisar los contextos de uso de las palabras. De acuerdo con Giammatteo y Albano (2009), consideramos ocho categorías posibles para la clasificación: preposiciones, pronombres, artículos, conjunciones, interjecciones, verbos, adjetivos y sustantivos.

Dado que las palabras escritas pueden tener naturalezas diferentes, al menos ser léxicas (o de contenido) o de función (o gramaticales), separamos las 14,174 palabras del corpus distinguiendo estas dos grandes categorías. En las Tablas 3 y 4 mostramos el resumen de frecuencias del número de ocurrencias (tokens) y tipos de

palabra (types) por categoría gramatical agrupando en la Tabla 3 las palabras léxicas y, en la 4, las de función[23].

Tabla 3.
Distribución de Palabras Léxicas en el Corpus General

Categoría Gramatical	Ocurrencias (tokens)	Tipos (types)
Sustantivos	35,109	5,549
Verbos	29,696	6,255
Adjetivos	23,665	3,001
Preposiciones	16,562	22
Adverbios	5,590	52
Total de palabras léxicas	**75,513**	**9,330**

Tabla 4.
Distribución de Palabras de Función (no léxicas) en el Corpus General

Categoría	Ocurrencias (Tokens)	Tipos (Types)
Interjección	2,527	39
Conjunción	13,661	29
Artículo	20,752	13
Pronombre	21,602	72
Onomatopeya	337	129
Total	**58,879**	**282**

Cabe señalar que en las tablas anteriores (3 y 4), la cifra que resulta del total de palabras no corresponde con el total de ocurrencias de palabras (tokens) ni con el total de tipos (types) del corpus general, debido a que hubo 934 palabras que por su uso dentro de los textos realizaban funciones gramaticales diferentes. Por ejemplo, “dulce”

[23]El Anexo 1 corresponde a las palabras enlistadas dentro de cada categoría gramatical acomodadas por frecuencia.

apareció como adjetivo (en la frase “el agua dulce”) y como sustantivo (en la frase “me compré un dulce”). En este caso, registramos “dulce” en su doble función[24].

Conocer la frecuencia de aparición de cada término y la categoría gramatical en la que se encuentra dentro del corpus general nos posibilitó describir las características de las palabras más frecuentes. A su vez, esta muestra del corpus general nos permitió inferir sobre el comportamiento de las palabras en el español.

Las 100 Palabras más Frecuentes

Para completar la descripción general del corpus, consideramos a las 100 palabras más frecuentes. Dentro de este conjunto, las palabras funcionales fueron las más frecuentes (63%) en comparación con las palabras léxicas (37%). La primacía de palabras funcionales coincide con los datos descriptivos de la lengua española reportados por Cantos y Sánchez (2011). La preeminencia de palabras de función se corrobora con los datos presentados respecto a la longitud de palabras. De esta manera no es coincidencia que las palabras de una y tres letras representen el 45% de la totalidad de palabras si consideramos que los artículos, las preposiciones y las conjunciones en su mayoría se conforman por este número de letras.

[24] En el Anexo 2 aparece el listado de las 207 palabras que se presentaron en el corpus como sustantivo y como otra función gramatical.

El Análisis de los Sustantivos

Dado que el propósito de este trabajo fue dar cuenta de los sustantivos escritos con mayor frecuencia en los textos dirigidos a un público infantil, después de realizar una caracterización del corpus general nos concentramos exclusivamente en las palabras sustantivas.

Clasificamos las 5,549 palabras que integran la categoría de sustantivos en dos subclases léxicas. Para ello, consideramos a los sustantivos que tuvieran la posibilidad de pertenecer a una clase como sustantivos comunes y aquellos que tienen la posibilidad de designar a una entidad como sustantivos propios. Decidimos subsecuentemente analizar solo los sustantivos comunes; es decir, 5,191. El resto (434 sustantivos propios) fueron excluidos ya que consideramos que rebasaban los propósitos de esta tesis. La Tabla 5 muestra la distribución de los sustantivos de acuerdo con su subclase léxica.

Tabla 5.
Frecuencia y Tipos de Sustantivos

Subclases Léxicas	Frecuencias (tokens)	Tipos de palabra (types)
Sustantivos propios	2,128	434
Sustantivos comunes	32,981	5,115
Total	**35,109**	**5,549**

Como se puede observar en la tabla anterior, el número de ocurrencias acumuladas de los sustantivos comunes fue de 32,981. Esta cifra corresponde al 43.55% de las palabras léxicas presentes en el corpus general, por lo que podemos afirmar que constituye una muestra representativa.

Creación del Vocabulario Fundamental

Nuestro análisis continuó con la lematización de los sustantivos comunes. Para ello, realizamos dos tareas simultáneas: la organización lemática de los términos y el establecimiento de frecuencias de aparición de dichos términos.

De acuerdo con la tradición lexicográfica (Lara, 2006), las cabezas léxicas se presentaron en la forma del sustantivo masculino y singular; por ejemplo, “abuelo” fue el lema para “abuelas”, “abuelitos” y “abuelito”. En algunos casos, dentro del corpus no apareció la forma convencionalizada de la cabeza léxica (por ejemplo, solo encontramos, “abejas”, y “abejita”). En estos casos creamos la cabeza léxica (siguiendo con el ejemplo, “abeja”) y bajo ésta agrupamos y cuantificamos las ocurrencias relativas. Así, en este mismo ejemplo sumamos la frecuencia de cada palabra “abejas” (26), “abejita” (2) bajo el vocablo “abeja” con 28 ocurrencias.[25]

Una vez que establecimos las cabezas léxicas (3,328 en total) procedimos con la determinación del vocabulario fundamental. Si bien habría sido suficiente seguir con el criterio propuesto por Chande (1979), utilizado para determinar el vocabulario fundamental del español de México, consideramos que analizar solo 44 lemas no resultaría útil para las aplicaciones psicolingüísticas a las que pretende beneficiar nuestro trabajo.[26]

[25] El Anexo 3 presenta el listado completo de cabezas léxicas junto con las palabras correspondientes al mismo paradigma de palabras. Se muestra también la ocurrencia acumulada de cada una.

[26] Bajo el criterio de Chande (1979) la inclusión de un lema como vocablo se da cuando la palabra se encontró dentro del primer cuartil (acumulación del 25% de las ocurrencias más frecuentes) de la lista total de lemas del corpus de sustantivos. En total, los vocablos considerados dentro de este primer cuartil fueron 44 lemas

El anexo 5 muestra las palabras del corpus general clasificadas por cuartiles, de forma especial se señalan los 44 lemas incluidos en el primer cuartil.

El Anexo 6 muestra las palabras incluidas en nuestro vocabulario fundamental.

Utilizamos, por lo tanto, el criterio empleado por el *Centre de Recherche et d' Etude pour la Diffusion du Francais* (para establecer el vocabulario fundamental del francés). Este grupo de investigación, para incluir un término considera exclusivamente su frecuencia de aparición. Así, aquellas palabras que tuvieron frecuencia igual o mayor a 20 ($\underline{F >}$) forman parte del vocabulario fundamental (Lara, 2006). Anexar este criterio para la selección de lemas nos permitió agregar 316 palabras a nuestro vocabulario fundamental. Con ello, incrementamos el porcentaje de palabras a analizar de un 1.32 %, que representaban los 44 lemas iniciales respecto al total de lemas, a un 10.81%. Por lo tanto, nuestro vocabulario fundamental constó de 360 vocablos. A partir de esta muestra realizamos el análisis central para los fines de esta tesis.

Longitud de palabras del vocabulario fundamental de acuerdo con el número de sílabas que las conforman

Como parte de la descripción general del vocabulario fundamental de nuestro corpus atendimos a la longitud de los lemas. Al respecto, es importante señalar que existen diversos modos de medir la longitud de una palabra: de acuerdo con el número de letras que la conforma, al número de morfemas que posee o al número de sílabas que contiene. Para describir a las palabras que conformaron nuestro vocabulario fundamental tomamos como parámetro de medición el número de sílabas que conforman una palabra. Así, consideramos cuatro categorías posibles: palabras monosílabas, bisílabas, trisílabas y tetrasílabas. En la Tabla 6 presentamos la

clasificación de los lemas del corpus de sustantivos frecuentes de acuerdo con el número de sílabas que contienen.[27]

Tabla 6.
Longitud Silábica de los Lemas del Vocabulario Fundamental por Número de Palabras y Porcentaje

Longitud de palabra	Número de palabras	Porcentaje
tetrasílabas	15	3.89
trisílabas	102	28.33
bisílabas	219	60.83
monosílabas	25	6.94
Total	**360**	**100.00**

Con estas cifras podemos concluir que en el corpus de sustantivos escritos más frecuentes predominaron las palabras bisílabas, con más de la mitad de palabras del total, seguidas por las palabras trisílabas con más del 20%, mientras que las palabras con 4 sílabas representaron el menor porcentaje (3.8%).

La distribución de vocablos de acuerdo con su longitud silábica fue también reportada en los datos estadísticos del vocabulario fundamental de México (Lara, 2007). Estos datos, confirmaron los resultados de diversos estudios en torno a las características de las sílabas en el español. [28] Estos estudios reportaron que la longitud de vocablos de acuerdo con su frecuencia en el español tiene un orden: las palabras bisílabas son las más frecuentes; en segundo lugar están las trisílabas; en tercera

[27] El anexo 4 muestra los lemas del vocabulario fundamental de acuerdo con su longitud silábica
[28] Las comparaciones se realizaron con las bases de datos: "The Statistical Properties of the Spanish Lexicon" Urrutibéhrity, 1972 (en Lara, 2007),"Los tipos silábicos del español", García, 1985 (en Lara, 2007) y "Las sílabas básicas del español según sus restricciones fonotácticas" Guirao y García, 1989 (en Lara, 2007).

posición se encuentran las palabras tetrasilábicas y, por último, las monosilábicas. Afirmaron también la presencia de estructuras CV y CVC como tipos silábicos predominantes en este idioma.

Los datos anteriores muestran que los resultados de nuestro estudio concuerdan nuevamente con los resultados del Vocabulario Fundamental de México. Concordamos en la predominancia de las palabras bisílabas y trisílabas dentro de nuestro vocabulario fundamental, más del 80% de las palabras se encuentran clasificadas en estas categorías y con la presencia de las estructuras CV y CCV como estructuras predominantes al estar presentes en un 72% de las sílabas de nuestro vocabulario fundamental (como lo mostraremos a continuación). Hasta aquí hemos mostrado primero las características del corpus general, después las características del vocabulario fundamental creado a partir de las palabras sustantivas. En lo que sigue, mostraremos los resultados del análisis de las sílabas iniciales de nuestro vocabulario fundamental.

Composición silábica del vocabulario fundamental de sustantivos

El análisis de estructuras silábicas presentes en los vocablos lo realizamos con ayuda del programa estadístico SPSS. Éste constó de dos fases. Primero identificamos los tipos de estructura silábica presentes (types) y su frecuencia de aparición (tokens). Después contabilizamos las posibles combinaciones de letras dentro de cada una.

En los 360 vocablos encontramos la presencia de 828 sílabas (tokens), de las cuales 295 son sílabas diferentes (types). La Tabla 7 contiene los 11 diferentes tipos de estructura silábica presentes bajo la siguiente nomenclatura:

CV Sílaba compuesta por una consonante y un núcleo vocálico; por ejemplo, *"ta"* en *"taza".*

CVC Sílaba compuesta por una consonante inicial, un núcleo vocálico y una consonante en la coda; por ejemplo, *"sol"* en *"soldado".*

CCV Sílaba compuesta por dos consonantes iniciales y un núcleo vocálico; por ejemplo, *"Tra"* en *"trapo".*

CVV Sílaba compuesta por una consonante inicial y dos vocales. Por ejemplo, *"gua"* en *"agua".*

CVVC Sílaba compuesta por una consonante inicial y otro final y dos vocales intermedias**.** Por ejemplo, *"pue"en "puerta".*

CCVC Sílaba compuesta por dos consonantes iniciales, un núcleo vocálico y una consonante en la coda. Por ejemplo, *"trac"* en *"tractor".*

CCVV Sílaba compuesta por dos consonantes iniciales y dos vocales finales. Por ejemplo, *"trai"* en *"traigo".*

CVCC Sílaba compuesta por una consonante inicial, seguida de una vocal y coda con dos consonantes. Por ejemplo, *"cons"* en *"construcción"*

V Sílaba con presencia exclusiva de vocal; por ejemplo, la primera "*a"* en "*agua"*.

VV Sílaba compuesta por dos vocales; por ejemplo, *"ai"* en *"aire".*

Tabla 7.
Frecuencia de Estructuras Silábicas en el Vocabulario Fundamental

Estructura silábica	Ocurrencia	Porcentaje
CV	514	62.15
CVC	110	13.3
CCV	47	5.68
CVV	44	5.32
V	39	4.72
CVVC	33	3.99
VC	26	3.14
CCVC	8	0.97
VV	4	0.48
CCVV	1	0.12
CVCC	1	0.12
Total	**827**	**100**

Como lo anunciábamos al final del apartado anterior, en la Tabla 7 podemos observar que la estructura CV predomina con 62.15%, seguida por la estructura CVC con 13.30%; juntas, representan más de la tercera parte de la totalidad de sílabas presentes en el corpus de sustantivos escritos más frecuentes (75.45%). Observamos también que las estructuras CCVC, VV, CCVV y CVCC fueron poco frecuentes, con una ocurrencia de 1 a 8, su acumulación representa el 1.69% del vocabulario fundamental.

Como ya lo mencionábamos, en el español, existe un patrón canónico de la sílaba que define las particularidades de nuestra lengua (Lara, 2006). Lara, Ham y García (1979), al elaborar el Vocabulario Fundamental de México, reportaron las características silábicas de nuestra lengua. Quisimos conocer si los resultados del análisis de las estructuras silábicas presentes en nuestro vocabulario fundamental correspondían con los datos reportados por estos investigadores por lo que

comparamos nuestros resultados con los del Vocabulario Fundamental de México. Así, constatamos que:

- Las estructuras más frecuentes coinciden. La presencia de las estructuras CV, CVC y VC en nuestro estudio representa el 78.59%. Este dato concuerda con los resultados del estudio del Vocabulario Fundamental de México donde se reporta que cerca del 75% de las sílabas en español están formadas por estructuras CV, CVC y VC.
- Las estructuras silábicas presentes en las palabras son muy similares. En nuestro vocabulario fundamental se presentaron 11 tipos de sílaba (types): CV, CVC, CCV, CVV, V, CVVC, VC, CCVC, VV, CCVV, CVCC. Los resultados del Vocabulario Fundamental de México (Lara, 2006) considera 14 tipos de sílabas (types). Diferimos en la presencia de 3 estructuras silábicas: VVC, VCC, CCVVC. Sin embargo, estas estructuras representan aproximadamente el 0.54% del total de sílabas presentes en el vocabulario fundamental del español de México, por lo que consideramos que no representan una discrepancia significativa.

Combinaciones de letras por estructura

En cuanto a las posibles combinaciones de letras dentro de cada estructura hallamos que las estructuras CVC y CV tuvieron la mayor variedad de letras con más del 28% de combinaciones diferentes Las estructuras CVV, CCV y CVVC presentaron un rango medio de variedad de combinaciones, oscilando entre el 8.47% y el 12.20%. En cambio, las estructuras VC, V y CCVC, CCVV, V, CVCC mostraron poca variedad. La Tabla 8 muestra los contextos silábicos (combinaciones de letras diferentes)

presentes en el vocabulario fundamental y su distribución de acuerdo con su estructura silábica.

Tabla 8. Contextos silábicos y Porcentaje, del Vocabulario Fundamental de Sustantivos, por Estructura Silábica

Estructura	Tipos de sílaba	Porcentaje
CV	86	29.15
CVC	83	28.14
CCV	25	8.47
CVV	36	12.20
V	9	3.05
CVVC	28	9.49
VC	17	5.76
CCVC	7	2.37
VV	2	0.68
CCVV	1	0.34
CVCC	1	0.34
TOTAL	**295**	**100.00**

Como lo hemos venido diciendo, el vocabulario fundamental de sustantivos se conformó de 360 palabras, en ellas encontramos que las sílabas iniciales caen en las 10 estructuras silábicas antes descritas para el Vocabulario Fundamental de México. Al mismo tiempo, se distribuyeron en 201 contextos silábicos diferentes[29]. El anexo 7 muestra el listado de sílabas presentes en el vocabulario fundamental y su estructura silábica, ordenadas de acuerdo con su frecuencia de aparición.

Como lo mencionamos en el Capítulo 2, la sílaba es definida como la estructura fundamental básica de toda agrupación de fonemas de una lengua dada. La separación silábica por tanto, es un fenómeno fundado en principios estrictamente fonológicos. De acuerdo con Obediente (1998), la división silábica se rige por un criterio funcional, que

[29] Entendemos por contexto silábico las diferentes combinaciones de letras que integran una sílaba. Por ejemplo, para la estructura CV podemos encontrar los contextos: la, pa, ma, ra, etc.

determina a un conjunto de sílabas de acuerdo con su estructura fonemática dependiendo de la lengua en que se encuentre. Así, por ejemplo, en la palabra “contigo” la frontera silábica no puede estar antes del grupo consonántico ya que la combinación “nt”, en español, no es posible como inicial de sílaba.

La división silábica de las palabras es, como lo señala la Real Academia Española (2010), un problema fonológico. Sin embargo, de acuerdo con la ortografía del español, la representación escrita de las palabras hace que, en algunos casos, un fonema sea representado con más de dos grafías y que estas, en ocasiones, no coincidan con lo que convencionalmente se designa como “consonante” o “vocal”. Por ejemplo, “chile” puede dividirse silábicamente en dos fragmentos: “chi-le” en donde ambas sílabas presentan la misma estructura CV a pesar de que en la primera sílaba la ortografía demanda el dígrafo “ch” para representar el fonema /č/. En este mismo sentido, la ortografía de algunas palabras en español demandan del uso de letras (vocales o consonantes) de las que se omite el valor sonoro; por ejemplo “u” en “que” o “gue” o bien “h” en “hijo” u “hoja”.

Por estos rasgos de ortografía irregular (en tanto que una grafía no necesariamente corresponde con un fonema), nombramos a este tipo de sílabas “especiales”.

Como nuestro trabajo pretendió dar cuenta de cómo son las palabras escritas, nos pareció relevante mencionar a las sílabas especiales, por lo que, al desglosar los contexto silábicos presentes en cada estructura señalaremos a estas sílabas.

Etiquetar a cada sílaba con la estructura silábica que la conformó nos permitió, primero, definirlas de manera general y, después, agruparlas y caracterizarlas con mayor detalle. A continuación, la Tabla 9 muestra la distribución de las 360 sílabas de acuerdo con su estructura silábica.

Tabla 9.
Estructuras Silábicas Presentes en Sílabas Iniciales por Ocurrencias y Tipos

Estructura	Ocurrencias (tokens)	Tipos(types)
CV	160	59
CVC	67	57
V	39	9
CVV	23	22
CVVC	22	21
VC	19	12
CCV	18	14
CCVC	8	7
VV	3	2
CVCC	1	1
TOTAL	**360**	**204**

Al analizar las estructuras silábicas presentes en las sílabas iniciales hallamos que, al igual que en el vocabulario fundamental, las estructuras CV y CVC se encontraron en la mayor parte de las sílabas iniciales (64.65%). En segunda posición encontramos a la estructura V con un porcentaje de 10.83%.

Las estructuras CVVC, CVV, CCV y CCVC estuvieron presentes en un menor porcentaje, oscilando entre un 2 y un 7%. Observamos también que las estructuras menos frecuentes fueron VV y CVCC con menos de 1%.

En cuanto a los contextos silábicos presentes en primera posición, la estructura CV presentó el mayor porcentaje de combinación de letras (92.85%) seguida por la estructura CVVC (91.30%). Con un porcentaje menor hallamos a las estructuras CVV, CCVC y CVC que presentaron más del 80% de variedad de letras. En cambio, las estructuras silábicas que presentaron un menor número de combinaciones fueron VV, CV, CVCC, CCV y V. Resulta interesante señalar que estas estructuras tienen poca variedad de letras debido a dos causas: 1) repetición de combinación de letras o, 2) minoría de sílabas iniciales con esta estructura, como en el caso de la estructura, CVCC donde hay una sola sílaba con esta estructura.

Las cifras anteriores fueron el resultado de valorar la lista general de sílabas iniciales[30]. En lo que sigue, desglosamos las sílabas que componen cada estructura silábica para conocer qué letras se hallaron en posición inicial y en qué combinaciones se presentaron. Mostraremos los resultados en tres apartados: Primero, mostraremos la composición de las tres estructuras silábicas más frecuentes que comienzan con una consonante. Después, analizaremos las estructuras silábicas que contienen una vocal en primera posición. Por último, describiremos las características de las estructuras silábicas que presentaron poca frecuencia dentro de nuestro corpus.

Estructuras silábicas frecuentes con una consonante como letra inicial

Dentro de las 10 estructuras silábicas presentes en nuestro corpus, siete de ellas contienen como letra inicial una consonante. De acuerdo con número de silabas contenidas en cada una, elegimos las tres más frecuentes para la descripción de sus

[30] El anexo 8 muestra el listado de sílabas iniciales ordenadas de acuerdo con su estructura silábica.

contextos silábicos: CV, CVC y CVVC. Al analizar las letras que componen las sílabas con estas estructuras, describimos al 69.44% del total de sílabas iniciales, por lo que resulta ser una descripción vasta y representativa del comportamiento de las consonantes como letras iniciales de palabra.

La estructura CV fue la más frecuente en las sílabas iniciales. El número de sílabas con esta estructura fue de 164, de las cuales 57 fueron diferentes, lo que indica que 34.75% del total de sílabas iniciales presentó contextos silábicos diferentes. Dentro de esta estructura encontramos como letras iniciales a 16 consonantes[31]. Ordenamos estas 16 letras por frecuencia de aparición y sacamos su distribución media. Así, seleccionamos a las letras que se encontraron por arriba de la media para la descripción de sus contextos. La Tabla 10 resume el porcentaje de ocurrencias de cada una de las principales consonantes iniciales.

Tabla 10.
Porcentaje de Presencia de Consonantes Iniciales más Frecuentes en Sílabas CV

Consonante inicial	Porcentaje de presencia
P	15.92
C	13.38
M	13.38
Total	**42.68**

En la tabla anterior, observamos que la consonante más frecuente fue “P” con 28 ocurrencias, estuvo presente como letra inicial en el 15.92% de sílabas, seguida por “C” y “M” con 21 apariciones (13.38%) cada una.

[31] El anexo 9 muestra las consonantes iniciales presentes en esta estructura junto con el porcentaje que representan.

La estructura CVC fue la segunda más frecuente en nuestro corpus al agrupar a 68 sílabas. También presentó un número alto de combinaciones de letras. De esta manera 57 sílabas de las 68 agrupadas en esta estructura fueron distintas, lo que representó un 83.82% de variabilidad. Las sílabas iniciales con estructura CVC, presentaron una variedad de iniciales de 13 letras. Comenzaron, sobre todo con “C”, “M”, “P”, y “S” en la siguiente proporción:

Tabla 11.
Porcentaje de Presencia de Consonantes Iniciales más Frecuentes en Sílabas CVC.

Consonante inicial	Porcentaje de presencia
C	19.12
M	14.71
S	13.24
P	11.99
Total	**40.13**

La aparición de las letras “M” y “P” como frecuentes coincide con las letras frecuentes de la estructura anterior (CV).Pero, en la estructura CVC, otra letra superó la media: la “S”.

La estructura CVVC presentó un porcentaje alto de variedad en la combinación de letras que la componen (91.30%). Dentro de las 23 sílabas categorizadas en esta estructura 21 fueron diferentes. En ellas, encontramos 10 consonantes como letras iniciales. Al igual que en la estructura CVV la letra más frecuente fue “C” (21.74%), seguida por las leras “D” y “M”. Cada una se presentó como letra inicial en el 13.04% de palabras. Las consonantes “V”, “T”, “S”, “P” y “F” ocuparon la tercera posición al distribuirse como letras iniciales, cada una en un 8.70% de palabras.

Al describir las tres estructuras más frecuentes que contienen a una consonante como letra inicial, observamos que las letras “C”, “P” y “M” se encontraron como frecuentes. A continuación, en las tablas 12, 13 y 14, describimos los contextos silábicos en los que se presentó cada una junto con su porcentaje.

Tabla 12.
Contextos Silábicos de la Consonante "C" en Posición Inicial por Estructura Silábica

Estructura silábica	Contextos silábicos					
	A	**E**	**I**	**O**	**U**	**Total**
CV	11	0	0	10	0	21
CVC	6	0	1	3	0	10
CVVC	0	0	1	0	4	5
Total	**17**	**0**	**2**	**13**	**4**	**36**

Tabla 13.
Contextos Silábicos de la Consonante "P" en Posición Inicial por Estructura Silábica

Estructura silábica	Contextos silábicos					
	A	**E**	**I**	**O**	**U**	**Total**
CV	14	5	5	4	0	28
CVC	4	3	0	0	1	8
CVVC	0	0	1	0	1	2
Total	**18**	**8**	**6**	**4**	**2**	**38**

Tabla 14.
Contextos Silábicos de la Consonante "M" por Estructura Silábica en Posición Inicial

Estructura silábica	Contextos silábicos					
	A	**E**	**I**	**O**	**U**	**Total**
CV	10	3	1	3	4	21
CVC	4	1	0	3	1	9
CVVC	0	0	1	0	2	3
Total	**14**	**4**	**2**	**6**	**7**	**33**

Las tablas anteriores nos permiten observar que el contexto silábico más frecuente fue la combinación de consonantes con la vocal “A”. La vocal “A” acompaña a una consonante en 49 ocasiones lo que representa el 45.79% de las sílabas. En segunda posición hallamos a la vocal “O”, con 23 apariciones (21.49%). En tercera posición, encontramos a la vocal “U” con 13 ocurrencias (12.14%). Por último, las vocales menos frecuentes fueron la “E” y la “I” con 11.21% y 9.34% respectivamente.

En cuanto a la presencia de las sílabas “especiales”, hallamos que dentro de la estructura CVC la letra “H” se presentó en una ocasión acompañando a la consonante “C”. Aunque su ocurrencia fue baja, nos pareció relevante señalarla debido a que se presentó acompañando a una consonante frecuente como inicial de palabra.

Estructuras silábicas frecuentes con una vocal inicial

Consideramos a las sílabas con estructura V y VC en posición inicial, para caracterizar los contextos silábicos en los que las vocales se presentan, por ser dichas estructuras las que mayor número de sílabas diferentes contienen (61 sílabas en total).

La estructura V se presentó en 39 sílabas iniciales. Encontramos entre ellas 9 diferentes realizaciones: las 5 vocales, más la combinación de ellas con la consonante “H” como letra inicial.[32] La vocal más frecuente fue “A” al estar presente en 48.72% de sílabas, seguida por “O” con un 23.08%. La vocal “E” ocupó el 4° lugar de frecuencia al presentarse en un 7.69% de sílabas. Por último, encontramos la presencia de las vocales “I” e “U” cada una presente en un 2.56% de sílabas. La Tabla 14 resume esta información.

[32] Contabilizamos dentro de la estructura “V” a las sílabas que inician con la consonante “H” y se acompañan de una vocal por la forma fonológica que presentan.

Tabla 15.
Ocurrencias y Tipos de Sílaba con Estructura V en Posición Inicial

Vocal Inicial	Ocurrencias(tokens)	Porcentaje
A	19	48.71
E	3	7.69
I	1	2.56
O	9	23.07
U	1	2.56
Total	**33**	**84.61**

El total de porcentaje en la tabla anterior no suma el 100% debido a que las sílabas que inician con la consonante “H” y se acompañan de una vocal se presentaron en 6 ocasiones por lo que conforman el 15.38% restante. Observamos que este porcentaje, ocupa el tercer lugar por frecuencia de aparición, por lo que aunque fonológicamente carezca de relevancia, la presencia de la letra “H” como inicial de palabra es un dato a considerar para el reconocimiento visual de palabras.

La estructura VC presentó una mayor combinación de letras. El 92.85% de las sílabas clasificadas en esta estructura fueron sílabas diferentes. Encontramos que la letra inicial más frecuente fue “E” al presentarse como letra inicial en el 45.35% de sílabas, en 36.35% como letra sola y en 9% antecedida de la letra “H” como inicial. En segunda posición, hallamos a las sílabas que inician con la consonante “H” precedida por las vocales “A”, “E”, “I” y “O” (27.30%). En tercera posición hallamos a la letra “A” con 18.2% seguida por “I” con 13.6%. La letra inicial menos frecuente en esta estructura fue “O” con 4.55% del total de sílabas.

En cuanto a los contextos silábicos, las letras con las que se combinaron las vocales fueron “S”, “H”, “M”, “N”, “R” y L”. Considerando la media de aparición de cada letra describiremos únicamente a los contextos silábicos de las letras más frecuentes:

“H” y “E”. Las Tablas (16 y 17) muestran los contextos silábicos que acompañaron a estas letras.

Tabla 16.
Contextos silábicos de “E” en estructura VC

Contextos silábicos	Ocurrencias	Porcentaje
M	1	4.55
S	7	31.82
Total	**8**	**36.37**

Tabla 17.
Contextos silábicos de “H” en estructura VC

Contextos silábicos	Ocurrencias	Porcentaje
A	1	4.55
E	2	9.09
I	1	4.55
O	1	4.55
Total	**5**	**22.74**

Describir los contextos silábicos anteriores es relevante porque las sílabas que contienen a las letras “E” y “H” conforman casi el 60% (59.11%) del total de sílabas VC en posición inicial.

La letra que mayor frecuencia presentó como contexto fue “S”, seguida por “M”. En el caso de las vocales que acompañan a la letra “H”, la “E” se presentó el doble de veces que el resto de las vocales. Es importante considerar, nuevamente, la frecuencia con la que esta letra se encontró dentro del vocabulario fundamental porque en el caso de la estructura VC, como mostramos en la tabla anterior, la letra “H” estuvo presente en el 22.74% de las sílabas.

Estructuras silábicas menos frecuentes

Como estructuras silábicas menos frecuentes encontramos a las siguientes estructuras: CVV, VV, CCV, CCVC, CCVC y CVCC. Dentro de la estructura CVV agrupamos 22 sílabas, de las cuales 20 fueron diferentes, lo que representó un 90.90%. Como letras iniciales en esta estructura hallamos a diez consonantes; las más frecuentes fueron “C” y “R” al estar presentes como letras iniciales, cada una en un 18.2% de palabras.

En la estructura CCV encontramos la presencia de 18 sílabas, de las cuales 14 fueron diferentes. Como primera letra encontramos a siete consonantes diferentes. Otra vez, como sucedió en la estructura CV y CVC, la consonante “P” fue la letra más frecuente al presentarse al inicio del 33.33% de las sílabas.

La estructura CCVC agrupó a ocho sílabas en total, en la cuales hubo siete sílabas distintas. Al analizar estas sílabas encontramos que tres inician con la consonante “P”, dos con “F”, dos con “T” y una con la consonante “G”. De nueva cuenta, la letra “P” predominó al presentarse en un 37.50% de las sílabas como letra inicial.

La estructura VV fue poco frecuente dentro de las sílabas iniciales. Esta estructura agrupó tres sílabas en las cuales encontramos dos tipos diferentes. Hallamos la combinación de la vocal “A” con la vocal “I” y la unión de la vocal “U” con la vocal “E” precedida por la consonante “H”.

La estructura CVCC fue la menos frecuente dentro de nuestro vocabulario fundamental, con una sola aparición.

Análisis general de sílabas en primera posición

Al analizar por cuartiles todas las letras que conformaron las sílabas iniciales, observamos que el primer cuartil se conformó por las consonantes “P” y “C”. En el segundo cuartil se agruparon las letras: “M”, “A”, “S”, “H”. Como era de esperarse en la distribución por cuartiles, a menor número de ocurrencia aumentaron los tipos de letras, por lo que después de la media encontramos 16 letras distribuidas entre el tercer y cuarto cuartil. Dentro del tercer cuartil hallamos a las letras “R”, “B”, “D”, “T”, “V”. En el último cuartil encontramos a F”, “G”. “E”, “L”, “N”, “O”, “I”, “J”, “Z” y a “U” ordenadas por su frecuencia de aparición. En esta distribución, es de notar que las consonantes “K”, “Ñ”, “X”, “Y” y “W” no se encontraron en ningún caso al inicio de sílaba. La Tabla 18 muestra las letras iniciales junto con el número de ocurrencias y porcentaje respecto a la totalidad de sílabas iniciales (360) ordenadas por cuartiles.

Tabla 18.
Letras Iniciales del Vocabulario Fundamental por Cuartiles

Cuartiles	Letra inicial	Ocurrencias (Tokens)	Frecuencia relativa	Frecuencia acumulada
	P	47	13.06	13.06
Uno	C	45	12.50	25.56
	M	36	10.00	35.56
	S	25	6.94	42.50
	A	24	6.67	49.17
Dos	H	20	5.56	54.72
	R	19	5.28	60.00
	B	17	4.72	64.72
	D	17	4.72	69.44
	T	17	4.72	74.17
Tres	V	15	4.17	78.33
	F	13	3.61	81.94
	G	12	3.33	85.28
	E	11	3.06	88.33
Cuatro	L	11	3.06	91.39

N	10	2.78	94.17
O	10	2.78	96.94
I	4	1.11	98.06
J	4	1.11	99.17
Z	2	0.56	99.72
U	1	0.28	100.00
Total	**360**	**100.00**	

A continuación, describimos las características fonológicas de las letras que se agruparon en los dos primeros cuartiles.

Primer cuartil

Las letras “P” y “C” fueron las letras iniciales más frecuentes. Ambas corresponden a fonemas oclusivos sordos (con excepción de “C” antes de “E” o “I”). De acuerdo con Alarcos (1965), su sonoridad es grave. Están clasificadas como fonemas simples porque, a diferencia de otras consonantes, no requieren de un articulador adicional para su pronunciación. De acuerdo con la hipótesis de explosión Davies, 1984 (en Dekker y Bob, 2006), era de esperarse que tuvieran una frecuencia alta debido a que, bajo esta teoría, las consonantes que requieren más de una acción articulatoria son más difíciles de producir y por ende serán menos frecuentes. Si fuera así, los resultados de nuestra tesis confirmarían la hipótesis.

Segundo cuartil

Las letras que encontramos en el segundo cuartil son: S, M, A y H. “M” y “S” coinciden en representar fonemas continuos. Al mismo tiempo, la letra “M” representa un fonema nasal mientras que “S” representa a otro fricativo estridente. La “A” representa un fonema vocálico clasificado como el de mayor percatabilidad en el

español por ser, de acuerdo con Gill (2007), una vocal abierta, central y anterior. La letra “H” aunque, como lo mencionamos anteriormente, no se encuentra caracterizada desde la fonología, estuvo presente en un número considerable de veces (20), por lo que obtuvo una posición relevante como letra inicial.

Comparación de los Vocablos Fundamentales con Otras Corpora Similares

Para corroborar los datos obtenidos en nuestro corpus realizamos varios comparativos con trabajos realizados para el español.

CREA

El corpus CREA es la base de datos elaborada a finales de los años 90 por la Real Academia Española. Contiene 154’279,050 formas que documenta a partir de dos fuentes: español oral (10%) y escrito (90%). Como base de datos utilizó libros, revistas y periódicos procedentes de España e Hispanoamérica realizados y distribuidos en contextos adultos.

Como parte del análisis de sus resultados, muestra las 1,000 palabras más frecuentes del español[33]. Nos pareció interesante conocer si las palabras que conformaron nuestro vocabulario fundamental se encontraban dentro de este listado.

Al comparar las listas de palabras, hallamos que 151 lemas de los 360 que conformaron nuestro vocabulario fundamental se encontraron presentes dentro de las palabras más frecuentes del español, lo que representa al 41.94%. De los 151 lemas coincidentes, encontramos 146 en su forma de vocablo (masculino y singular) y 5 con alguna palabra perteneciente al mismo paradigma de palabras. Por ejemplo, la palabra

[33] Para consultar las 1000 palabras más frecuentes del español DEL CREA O DEL revise el anexo 1

"animales" se encuentra incluida dentro del vocablo "animal". En estos casos, contabilizamos a esta familia léxica como coincidente con las palabras más frecuentes del listado del CREA. A continuación, enlistamos las palabras coincidentes ordenadas por orden alfabético; junto se muestra el lugar de aparición que presentó dentro de la base de datos del CREA. La numeración es ordinal, comienza con el número 1 que asigna a la palabra más frecuente, la numeración es ascendente pero ordena a las palabras de mayor a menor frecuencia de acuerdo con número de ocurrencias que presentaron dentro del corpus.

Tabla 19 A
Palabras coincidentes entre el CREA y el Vocabulario Fundamental de este trabajo junto con su lugar de aparición dentro del CREA, por orden alfabético.

Palabra	Lugar de aparición	Palabra	Lugar de aparición	Palabra	Lugar de aparición	Palabra	Lugar de aparición
Agua	207	Don	281	Lado	233	Pasado	158
Aire	490	Dos	41	Libro	354	Paso	326
Amigo	691	Ejemplo	215	Llama	924	Persona	357
Animal	974	Era	45	Llamado	827	Pie	674
Año	102	Escuela	719	Lugar	139	Piel	966
Atención	484	Espacio	424	Madre	255	Poder	166
Ayuda	636	Especie	704	Mal	316	Pregunta	862
Bien	69	Estado	104	Manera	190	Problema	280
Boca	672	Este	33	Mano	238	Pueblo	277
Cabeza	291	Falta	308	Mañana	272	Puerta	374
Calle	378	Familia	30	Mar	558	Punto	201
Cama	895	Fin	186	Medio	160	Realidad	231
Cambio	236	*Flor*	992	Mes	472	Respuesta	587
Camino	406	Fondo	409	Mesa	611	Rey	597
Campo	384	Forma	107	Metro	729	Río	708
Cara	457	Frente	182	Miedo	743	Rosa	961
Carne	923	Fuego	806	Mil	328	Saber	387
Carta	740	Fuerza	368	*Millón*	132	Segundo	297
Casa	122	Gente	247	Modo	257	Seguro	710
Cerca	481	*Gracia*	542	Momento	131	Seis	407
Cien	400	Grupo	173	Muerte	264	Semana	321
Cinco	259	Hecho	125	Muerto	929	Señor	327
Ciudad	187	Hermano	932	Mujer	181	Siete	656
Clase	602	Hijo	398	Mundo	101	Siguiente	948
Color	609	Historia	194	Música	381	Sobre	32
Compañero	690	Hombre	135	Nada	109	Sol	567
Corazón	734	Hora	275	Negro	753	Son	40
Cosa	312	Idea	414	Niño	467	Suelo	649
Cuarto	720	Joven	469	Noche	205	Sueño	848

Cuatro	198	Juez	774	Nombre	265	Suerte	824
Cuenta	174	Don	281	Ocho	715	Tarde	225
Cuerpo	250	Dos	41	*Ojo*	263	Tiempo	70
Deseo	945	Ejemplo	215	Oro	860	Tierra	295
Día	90	Era	45	Padre	246	Trabajo	142
Dicho	299	Escuela	719	País	98	Tres	93
Dinero	394	Espacio	424	Palabra	554	Uno	67
Dios	369	Especie	704	Papel	411	Verdad	285
Doctor	764	Estado	104	Parte	64	Vez	58

Nota: Las palabras en itálica indican los lemas que no se encontraron en forma de vocablo

Tabla. 19b.
Últimas 7 palabras coincidentes entre el CREA y el Vocabulario Fundamental de este trabajo junto con su lugar de aparición dentro del CREA, por orden alfabético

Palabra	**Lugar de aparición**
Viaje	713
Victoria	905
Vida	76
Viejo	651
Vino	766
Voz	325
Vuelta	817

Corpus CUMBRE

Además del CREA, dentro de las corpora de referencia hechas para contextos hispanoparlantes, encontramos al corpus CUMBRE. Es un corpus de referencia del español contemporáneo elaborado por Aquilino Sánchez de la Universidad de Madrid. Está conformado por 20 millones de palabras que representan al español oral y escrito de España e Hispanoamérica a partir de considerar 3 fuentes: textos extraídos de libros de diversos géneros, textos extraídos de la prensa (secciones de diarios y revistas) y textos orales, procedentes de radio y televisión. [34] Nuevamente, la base de datos fue

[34] Para conocer la metodología de este corpus, consultar a Sánchez, A., Sarmiento, R., Cantos, P., y Simón, J. (1995). *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos y aplicaciones*. Madrid: SGEL MEJOR A TEXTO Y A BIBLIOGRAFÍA

tomada a partir de textos producidos para contextos adultos. A partir de las palabras que recopila este corpus, se elaboró una lista de palabras frecuentes etiquetadas por categoría gramatical. Buscamos las palabras que se presentaron dentro de la categoría de sustantivos, 35 en total, en los lemas de nuestro vocabulario fundamental. Encontramos que 26 de ellas se encuentran en nuestro listado, lo que representa que consideramos en nuestro análisis al 74.28% de sustantivos frecuentes del castellano, de acuerdo con esta base de datos. La Tabla 20 muestra los sustantivos del vocabulario fundamental coincidentes con los sustantivos frecuentes del CUMBRE.

Tabla.20
Palabras coincidentes entre las palabras frecuentes del Corpus CUMBRE y el Vocabulario Fundamental de este trabajo, por orden alfabético.

Lema	
Año	mujer
Casa	mundo
Cosa	niño
Día	noche
Don	ojo
Estado	padre
Forma	parte
Gente	persona
Grupo	punto
hombre	señor
Hora	tiempo
Mano	trabajo
momento	vez

Otro comparativo que realizamos de nuestros resultados, fue con el corpus SUBTLEX elaborado por la Universidad de Oviedo (España) en conjunto con la Universidad Ghent (Bélgica)[35]. Se trata de un corpus de referencia, 41 millones de palabras, realizado a partir de los subtítulos de series y películas en español. El objetivo de su construcción fue dar cuenta de las palabras más frecuentes en el español para realizar estudios relacionados con el reconocimiento de palabras. Como parte de sus resultados presentaron una lista de las palabras más frecuentes en nuestro idioma. Buscamos en ella, las palabras que conformaron nuestro vocabulario fundamental. Encontramos, que el 100% está enlistado. El anexo 10 muestra las

[35] Para conocer más sobre este corpus consultar, SUBTLEX-ESP: Spanish word frequencies en http://www.redalyc.org/articulo.oa?id=16920109001.

palabras más frecuentes del español de acuerdo a este estudio, en color azul marcamos los lemas que conforman nuestro vocabulario fundamental.

CAPÍTULO 5

CONCLUSIONES Y CONSIDERACIONES FINALES

La presente investigación consistió en la elaboración de un corpus de palabras escritas más frecuentes para niños. Para realizar este instrumento metodológico fue necesario indagar a través de la lexicografía el concepto de corpus y los requerimientos necesarios para su elaboración, de forma que deslindáramos esta herramienta de listados de palabras y otro tipo de recopilaciones lingüísticas.

Revisar cómo se elabora un corpus nos permitió entender que las corpora son herramientas que pueden utilizarse con diversos fines. Observar los datos a través de la lingüística descriptiva nos permitió redimensionar el estudio del léxico a través principalmente de la variable frecuencia de palabra. Descubrimos que más allá de un número de repeticiones, la frecuencia de una palabra refleja el índice de uso de los conceptos en nuestra cultura.

La comparación de los resultados del corpus general con otras corpora elaboradas para contextos hispanoparlantes nos permitió corroborar que nuestro corpus es representativo del español escrito. En primer lugar, observamos la similitud en cuanto a la **frecuencia de aparición de palabras**. De acuerdo con Torruella y Listerri (1999), la lista de unidades léxicas de un corpus, no importa qué tan grande sea, presenta un número elevado de palabras con una sola aparición (hápax legomena); en nuestro corpus, el 50% de los sustantivos del corpus fundamental coincide con las principales corpora de nuestra lengua. Los estudios del español

muestran también que en nuestro idioma, dentro de las palabras frecuentes, las palabras de función predominan sobre las palabras léxicas. Al analizar las primeras 100 palabras de nuestro corpus encontramos que 63% fueron palabras de función, con lo que corroboramos también este dato.

Concordamos también con la **longitud de palabras** que se reporta para el español. El estudio de Torruella y Listerri (1999) analiza la media de palabra en base a los datos del corpus CUMBRE Estos investigadores reportaron que la media de longitud de palabra es de 4.42 letras por palabra. En nuestro corpus la media de longitud de palabra fue 4.

Por otra parte, los datos del Español Fundamental de México (Lara, 2006) reportaron que las palabras en español se presentan con una frecuencia determinada de acuerdo con el número de sílabas que contienen. Como lo mostramos en el Capítulo 4, la distribución de nuestro vocabulario fundamental concordó con los datos presentados en esta investigación.

Observamos también que **las estructuras silábicas** que conformaron las palabras de nuestro corpus fueron similares a las reportadas en otras corpora del español. Las palabras de nuestro corpus contuvieron 11 estructuras silábicas diferentes. En los datos del Español Fundamental de México (Lara, 2006) se reportaron catorce tipos de estructuras silábicas. Al comprar las estructuras silábicas de ambas corpora observamos que las diez estructuras silábicas más frecuentes se encontraron en nuestros resultados y que las estructuras que no fueron equivalentes presentaron una frecuencia baja. Respecto a los sustantivos, es interesante advertir que al igual que

en otros estudios como el de Cantos y Sánchez (2011), éstos aparecen en posiciones retrasadas. El primer sustantivo en nuestro corpus se encontró en la posición número 37 . Todas las palabras anteriores fueron palabras de función o palabras que pueden clasificarse dentro de dos categorías gramaticales, como por ejemplo la palabra "como". La preeminencia de las palabras de función y de los verbos dentro de las primeras posiciones, de igual forma, correspondió con las caracterizaciones que se han realizado del español. Otro dato coincidente respecto a las estructuras silábicas fue la preminencia de las estructuras CV y CVC como las estructuras más frecuentes.

Comparar nuestro corpus con otras corpora realizados en español, nos permitió además de encontrar similitudes hallar discrepancias. Así, encontramos que las palabras frecuentes no son iguales en todos las corpora. Al comparar las palabras que conformaron nuestro vocabulario fundamental de sustantivos con las palabras más frecuentes de las corpora CUMBRE y CREA encontramos que no todas se encontraron enlistadas, solo el 74.28 y 41.94% respectivamente. Sin embargo, al buscar nuestros lemas dentro de los resultados del SUBTLEX hallamos que el 100% de éstas se enlistan.

Saber que los datos entre corpora no son exactamente iguales justifica la elaboración de nuevas corpora para continuar el estudio de una lengua. Elaborar un corpus especializado nos permitió en este estudio observar características de las palabras escritas para un público en particular (público infantil) que en corpora generales (de referencia) no es posible observar en su totalidad.

Es importante mencionar además que la tipología de corpus influye en la descripción que se haga de una lengua,. Así, los estudios comparativos entre palabras más frecuentes muestran que los corpus elaborados a partir de textos orales o mixtos son distintos a los que toman con base de datos textos escritos.

A lo largo de este trabajo hemos querido mostrar la importancia de contar con bases de datos que reflejen los usos escritos de las palabras y la frecuencia con las que éstas se presentan, sobre todo la utilidad que ha tenido para la psicolingüística. Al analizar los contextos silábicos presentes en el vocabulario fundamental encontramos algunos datos que apoyan resultados de estudios psicolingüísticos antes realizados.

Las tres letras más frecuentes como inicial de palabra fueron “P”, “C” y M”. Las tres son letras que corresponden a fonemas oclusivos (/p/, /k/, /m/), dos sordos y uno sonoro lo que apoyaría la hipótesis de articulación.

Los estudios de Cano y Vernon (2008) en torno al uso de los niños de las consonantes en diferentes contextos reportaron que las letras que tienen sonoridad continua como “M” y “S” parecen ser más fácilmente aislables y reconocibles que otras consonantes. La letra “M” fue una letra frecuente, lo que podría influir en la facilidad que existe para su reconocimiento. Estas investigadoras también reportaron que existe dificultad en la lectura de palabras cuando inician con las letras “I” o “U”. De acuerdo con nuestros resultados los contextos silábicos más frecuentes son “A” y “O”. Pensar que las letras “U” e “I” son poco frecuentes, podría apuntar a que la dificultad que presentan los niños al leer estas letras esté relacionada con el número de veces que éstas se encuentran como contexto silábico.

En este mismo estudio las investigadoras reportaron que las consonantes “P” y “M” presentaron la misma probabilidad de ser elegidas por los niños para la escritura de palabras. Plantearon como parte de sus resultados la hipótesis de la posible influencia que referentes como las palabras “papá” y “mamá” tienen en la selección de letras que realiza un niño. Sin embargo, en nuestros resultados la frecuencia que estas dos letras presentan es similar, por lo que probablemente sea la frecuencia de aparición en los textos la que influya en la selección de letras y no referentes específicos.

La utilidad del presente estudio reside en que proporciona datos sobre cuáles son los sustantivos escritos más frecuentes, con qué frecuencia de aparición se presentan las letras como inicial de palabra y en qué contextos silábicos están dentro de textos dirigidos a un público infantil. Nuestro trabajo constituye por tanto una herramienta metodológica que permitirá realizar estudios relacionados con diferentes variables: familiaridad léxica, familiaridad subjetiva de palabras, ambigüedad léxica (homonimia y polisemia), categoría gramatical, umbral de activación de letras, tomando en cuenta un universo distinto al que la gran mayoría de corpus realizados para contextos hispanoparlantes considera: palabras dirigidas para un contexto infantil.

Existe la necesidad de corroborar la trascendencia que estos datos pudieran tener para estudiar fenómenos ligados con selección de palabras más o menos frecuentes. Esta investigación abre posibilidades para el estudio próximo de otros términos gramaticales utilizando la misma base de datos ya procesada por el analizador léxico WST. De esta manera sabemos que no solo es posible analizar los sustantivos como lo hicimos en este trabajo, sino que es posible también realizar el estudio de palabras con diferentes categorías, determinar los contextos en los que

éstas se presentan e incluso dar cuenta de la relación entre palabras dentro de frases completas.

BIBLIOGRAFÍA

Alarcos, L. (1986). *Fonología Española*. Madrid: Editorial Gredos.

Alameda, J., y Cuetos, F. (1995*). Diccionario de las unidades lingüísticas del castellano Vol 1: Orden Alfabético, Vol II: Orden de frecuencias.* Servicio de publicaciones de la Universidad de Oviedo.

Alvarado, M. (1997). *Conciencia fonológica y escritura de niños preescolares: la posibilidad de omitir el primer segmento.* Tesis de maestría. Universidad Autónoma de México. Querétaro.

Anglada, E., y Bargalló, M. (2007). *Principios de lexicografía moderna en diccionarios del siglo XIX*. Alicante: Biblioteca Virtual Miguel de Cervantes.

Ávila, R. (1991, junio). *Densidad léxica y adquisición del vocabulario: niños y adultos en el español de América*. Actas del III Congreso Internacional del español de América, Valladolid, C. Hernández, G. P. Granada, et al. (eds.), Valladolid, 1991.

Ávila, R. (2006) *VALIDE: Variación léxica internacional del* español (programa de cómputo, idea y diseño). Instituto de Ingeniería UNAM. México: COLMEX

Cabré, M., y Bach, C. (2004). *El Corpus Tenic de IULA: Corpus textual especializado plurilingüe"* Panace@ - *Boletín de Medicina y Traducción* V(16): MedTrad. p. 173-176. Consultado en: http://www.medtrad.org/panacea/PanaceaPDFs/Panacea16_Junio2004.pdf

Cabré, M., y Bach, C. (2004) *Corpus IULA* (Corpus especializado plurilingüe). Revisado en www.iula.upf.edu/corpus/corpuses.htm

Cabré, M. (2007). *Constituir un corpus de textos de especialidad: condiciones y posibilidades".* En Ballard, M.; Pineira-Tresmontant, C. (ed.). *Les corpus en linguistique et en traductologie*. Arras: Artois Presses Université. 89-106. ISBN 978-2-84832-063-2

Calderón, G. (2010). La hipótesis alfabética y la conciencia fonológica en niños de preescolar. En Calderón y Hess *El Reto de la Lengua Escrita en la Escuela (45-66).* México:FUNDAP

Cano, S., y Vernon, S. (2008, septiembre*). Denominación y uso de consonantes en el proceso inicial de alfabetización*. Lectura y Vida. Revista Latinoamericana de Lectura*, 2 (29), 32-45.*

Cantos, P. y Sánchez, A. (2011). Corpus CUMBRE. Consultado en: *http://www.um.es/lacell/miembros/asp/textos/Cumb-cor.htm*

Chall, J. (1983). *Stages of Redding development*. New York: McGraw-Hill.

CHILDES. (2003, Octubre 18). *Childes Data Exchange System. Consultado el 30 noviembre 2013 en:* http://childes.psy.cmu.edu/

Cuetos, F. (2010). *Psicología de la lectura*. Wolters Kluwer España: Madrid.

Cuetos, F., González, M., Barbón, A., y Brysbaert, M. (2011). *SUBTLEX-ESP: Spanish word frequencies based on film subtitles.* Revista de metodología y psicología experimental, 32(2), p. 133-143. Consultado en: http://hdl.handle.net/10651/6265

Dávalos, A. (2008). *La puntuación en la organización de textos infantiles propios y ajenos*. Tesis de Maestría. Universidad Autónoma de Querétaro.

Davies, M. (2002, Diciembre 11) *Corpus del español Brigham Young University*. Consultado el 12 de Agosto 2012 en: http://corpus.byu.edu/bnc/

Davies, M. (2006). A Frequency Dictionary of Spanish: Core Vocabulary for learners. New York & London: Routledge.

Ehri, L.C. (1992). *Reconceptualizing the development of sight Word Redding and its relation to recoding*. En Gough, Eahri and Treiman (eds.) Reading acquisition, p. 107-145, Mahwah, NJ: Lawrence Earlbaum Associates.

Faber, P., Moreno, A y Pérez, Ch. (1999). *Lexicografía Computacional y Lexicografía de Corpus*. Volumen monográfico. RESLA. p. 175-213. Consultado en: http://lexicon.ugr.es/faber

Ferreiro, E. (1982) *Los procesos constructivos de apropiación de la escritura* en: E. Ferreiro y M. Gómez Palacio, *Nuevas perspectivas sobre los procesos de lectura y escritura*. México: Siglo XXI.

Ferreiro, E. y Teberosky, A. (1979). *Los sistemas de escritura en el desarrollo del niño*. México: Siglo XXI.

Ferreiro, E. y Zamudio, C. (2008). *La escritura de sílabas CVC y CCV en los inícios de la alfabetización escolar. ¿Es la omisión de consonantes prueba de incapacidad para analizar La secuencia fônica?* Rivista di Psicolinguistica Applicata, VIII, 1-2, 37-53

Forster, K y Forster, J. (2003). *DMDX: A Windows display program with millisecond accuracy*. University of Arizona, Tucson, Arizona Behavior Research Methods, Instruments, & Computers. 35(1), 116-124. Consultado en: http://www.indiana.edu/~clcl/Q550_WWW/Papers/ForsterEtAl(2003).pdf

Giammatteo, M. y Albano, H. (2009) *¿Cómo se clasifican las palabras?* Buenos Aires: Biblos.

Goswami, U. (2010). *Phonology, reading and reading difficulties.* En K. Hall, U. Goswami, C. Harrison, S. Ellis and J. Soler (eds.) *Interdisciplinary perspectives on lerning to read*. Culture cognition and pedagogy. New York: Routledge.

Hernández, R., Fernández, C y Baptista, P. (2007). Capítulo 11. El reporte de resultados del proceso cuantitativo. Capítulo 16. El reporte de resultados del proceso cualitativo. En R. Hernández Sampieri; C. Fernández Collado y P. Baptista, L. *Metodología de la Investigación* (pp.501-518 / pp.721-750).México: Mc. Graw Hill.

Hopkins, K., R. Hopkins y G, Glass. (1997). *Estadística Básica para las Ciencias Sociales y del Comportamiento*. México: Prentice-Hall Hispanoamericana.

Lara, L. (1996). *Teoría del diccionario monolingüe*. México: El Colegio de México, Centro de Estudios Lingüísticos y Literarios. ISBN 968-12-0705-X

Lara, L. (2005). *Diccionario del español usual en México*. México D.F.:El Colegio de México, Centro de Estudios Linguísticos y Literarios, c1996. ISBN 968-12-0704-1

Lara, L. (2006). *Curso de Lexicografía*. México: COLMEX

Lara, L. (2007). *Resultados numéricos del vocabulario fundamental del español de México*. México: El Colegio de México.

Lara, L., García, I y Ham, R. (1980). *Investigaciones lingüísticas en lexicografía*. México: COLMEX.

López, C., Faber P., y Tercedor, M. (2006). *Terminología basada en el conocimiento para la traducción y la divulgación médicas: el caso de Oncoterm*. Panace@ 7(24), 228-240. Extraído de: http://www.webcitation.org/5u3v2ao0g

Lozano, C. (2009). *CEDEL2: Corpus Escrito del Español L2*. En: Callejas B., Carmen, M. (eds.) Applied Linguistics Now: Understanding Language and Mind / La Lingüística Aplicada Hoy: Comprendiendo el Lenguaje y la Mente. Almería: Universidad de Granada. Extraído de:http://www.uam.es/proyectosinv/woslac/DOCUMENTS/Presentations%20and%20articles/LOZANO%20CEDEL2%20AESLA%20Almeria.pd7

Luque, J. (2004). *Aspectos universales y particulares de las lenguas del mundo*. Estudios de lingüística en español. Vol. 21. Consultado en: http://www.raco.cat/index.php/Elies/search/titles?searchPage=10

Marín, F. (1992). *El Corpus Oral de Referencia de la Lengua Española Contemporánea.* Informe del proyecto. Madrid: España. Consultado en: ftp://ftp.lllf.uam.es/pub/corpus/oral.

Mac Whinney, B (1995). *The CHILDES Project tool for analyzing talk*. Hilldale, NJ: Erlbaun.

McEnery, T y Wilson, A. (1996). *Corpus Linguistics.* Edinburgh:Edinburgh University Press. ISBN 0-7486-0808-7

Mc Enery, A., Xiao, R. (2005). *Character Encoding in Corpus Construction. En Developing Linguistic Corpora: a Guide to Good Practice* (ed. M. Wynne. Oxford: Oxbow Books. pp.47-58). Extraído de: http://ahds.ac.uk/linguistic-corpora

Moreno, Torre, Curto y de la Torre. (2006). *Inventario de frecuencia fonémicas y silábicas del castellano espontáneo y escrito*; Memorias de las IV Jornadas en Tecnología del Habla; Zaragoza 8 al 10 de noviembre.

O'Donnell, M. (2008, abril). *The UAM CorpusTool: Software for corpus annotation and exploration*. Proceedings of the XXVI Congreso de AESLA, Almeria, Spain. Consultado en: http://www.uam.es/proyectosinv/woslac/DOCUMENTS/Presentations%20and%20articles/ODonnellAESLA08.pdf

Obediente, E. (1998). *Fonética y fonología*. Universidad de los Andes Consejo de Publicaciones, Mérida.

Parodi, G. (2008). *Lingüística de Corpus: Una introducción al ámbito lingüística de corpus*. Revista de lingüística teórica y aplicada, 46(1), 93-120. Extraído de:http://dialnet.unirioja.es/servlet/articulo? codigo=2714916. ISSN 0033-698X.

Parkes, M. (1992). *Pause and effect: An introduction to the history of punctuation in the west*. Cambridge: Scolar Press.

Pérez, Ch. (2002). *Explotación de las corpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento*. Universidad de Málaga. (Vol.18).ISSN: 1139-8736. Extraído de: http://elies.rediris.es/elies18/

Pérez, M., Alameda, J. y Cuetos, F. (2003). *Frecuencia, longitud y vecindad ortográfica de las palabras de 3 a 16 letras del Diccionario de la Lengua Española* (RAE, 1992). *Revista Española de Metodología Aplicada*, *8*, 1-20. Revisado en: http://www.psico.uniovi.es/REMA/v8n2/a1/

Prieto, S., Mosqueira, E., & Vázquez, N. (2009). *Córpora y enseñanza de lenguas: Se buscan colocaciones*. I International Corpus Linguistics Conference (pp. 366–373). Murcia, Spain: Universidad de Murcia.

Quereda, L. y Rodríguez. (2006). *Microconcord*. Oxford University Press. Departamento de filología inglesa. Universidad de Granada. Extraído de: http://www.ugr.es/~lquereda/microconcord.htm

Quinteros, G. (1997). *El uso y función de las letras en el periodo prealfabético*. México: CINVESTAV.

Real Academia Española (RAE). (2011). *Ortografía de la lengua española. México: España.*

Rojo, G. (2008, agosto) *Lingüística de corpus y lingüística del español. Universidad de Santiago de Compostela* [Ponencia plenaria en el XV Congreso de la ALFAL .Montevideo. Edición electrónica en las actas del congreso (ISBN 978-9974-8002-6-7)]

Samaja, J. (2007). *Fase 4. Diseño de los procedimientos*. Buenos Aires, Editorial Eudeba.

Sampieri, R., Fernández, C., Baptista L. (2010). *Metodología de la Investigación*. México: Mc Graw Hill

Scott, M. (2006).Chapter 7.*Corpora and Leanguage Teaching*. En Origin an history of corpus. Extraído de: http://www.lexically.net/wordsmith/corpus_linguistics_links/Roemer%20208%20HSK%20CL%20chapter%20final%20print%20version.pdf

Sebastián, N., Martí, M.A., Carreiras, M. y Cuetos, F. (2000). *LEXESP. Léxico informatizado del español*. Barcelona: Universi de Barcelona

Senso, J., Magaña, P., Faber P y Vila, A. (2007,noviembre). *Metodología para la estructuración del conocimiento de una disciplina: el caso de PuertoTerm*. El profesional de la información, 16(6), pp.591-604.DOI: 10.3145/epi.2007.nov.06 Extraído de: http://www.academia.edu/1376322/Metodologia_para_la_estructuracion_del_conocimiento_de_una_disciplina_el_caso_de_PuertoTerm_1

Serra, M., Serrat, E., Solé, R.,Bel, A., y Aparici, M. (2000) . *La adquisición del lenguaje*. España: Editorial Ariel

Sinclair, J. (2005). *Corpus and Text - Basic Principles en: Developing Linguistic Corpora: a Guide to Good Practice,* (ed.) M. Wynne. Oxford: Oxbow Books: 1-16. Extraído de: http://ahds.ac.uk/linguistic-corpora/

Smith, F. (2005). Comprensión de la lectura: análisis psicolingüístico de la lectura y su aprendizaje. México: Trillas.

Teberosky, A. (1990). Re-escribiendo noticias: Una aproximación a los textos de adultos y niños en proceso de alfabetización. Anuario de psicología 47 (4)

The British National Corpus. (2007, enero 26). BNC XML Edition version 3. Oxford University Computing Services on behalf of the BNC Consultado en: http://www.natcorp.ox.ac.uk/

The ICE Project. (2009, diciembre). International Corpus of English. Consultado en: http://ice-corpora.net/ice/publics.htm

Torruella, J., Llisterri, J. (1999).Diseño de corpus textual y oral. En Blecua J., Clavería, M., Sánchez, G. Torruella, J. (Eds.).Filología e Informática. Nuevas tecnologías en los estudios filológicos. Barcelona: Seminario de filología e informática, departamento de Filología Española, Universidad de Barcelona-Editorial Milenio pp.45-75. Consultado en: http://liceu.uab.es/~joaquim/publicacions/Torruella_Llisterri_99.pdf

Tribble, C. (1997).Improvising Corpora for ELT: Quick-and-dirty Ways of Developing Corpora for Language Teaching. En: Lewandowska-Tomaszczyk, Barbara/Melia, Patrick J. (eds.), Practical Applications in Language Corpora. The Proceedings of PALC '97. Lodz: Lodz UniversityPress. Consultado en: http://www.ctribble.co.uk/text/Palc.htm

Uria, L., Hulden, M., Etxeberria, I y Alegría, I. (2011).Proceedings of the Workshop on Iberian Cross-Language Natural Language Processing Tasks. ICL. Consultado en: http://ceur-ws.org/Vol-824/paper10.pdf

Ruiz, A y Ueda, H. (2003). VARILEX, Variación léxica del español en el mundo, Proyecto internacional de investigación léxica. Pautas y pistas en el estudio del léxico hispano (americano) / coord. por Gerd Wotjak, 2003, ISBN 84-8489-083-X , págs. 141-278

Vargas, Ch.(2009). Instrucciones de uso de WordSmith Tools (v. 5.0) Departamento Filología Inglesa. Universidad de Alicante. Extraído de: rua.ua.es/dspace/bitstream/10045/3923/9/ManualWST4.pdf

Vernon, S (1989). *El proceso de construcción de la correspondencia sonora en la escritura (en la transición entre los periodos pre-silábicos y el silábico).* Tesis de Maestría dirigida por Emilia Ferreiro; DIE-CINVESTAV, México.

Vernon, S. (2002). Children's Analysis of Oral and Written Words. En Brockmeier, Wang y Olson. *Litercy, Narrative and Culture*. Curzon:NY. 229-244

Vernon,S y Calderón,G. (1999) *Letras y sonidos en la alfabetización inicial.* Cuadernos de trabajo Sistema de investigación Miguel Hidalgo. México: CONACYT

Zamudio, C. (2010). *Las consecuencias de la escritura alfabética en la teoría lingüística.* México: El Colegio de México Centro de Estudios Lingüísticos y Literarios. México: COLMEX.