UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE FILOSOFÍA Y LETRAS
COLEGIO DE LETRAS HISPÁNICAS

ELABORACIÓN DE UN CORPUS ETIQUETADO DE DISCURSO INFANTIL ESCRITO

T E S I S
QUE, PARA OBTENER EL TÍTULO DE
LICENCIADA EN LENGUA Y LITERATURAS HISPÁNICAS,
PRESENTA

SANDRA NAYELY RICHER MONROY

ASESORES: DR. GERARDO EUGENIO SIERRA MARTÍNEZ
DRA. CELIA DÍAZ ARGÜERO

CIUDAD UNIVERSITARIA, 2010

Índice

1. INTRODUCCIÓN

1.1. Antecedentes

Esta tesis se enmarca dentro del proyecto *Aprender mientras se enseña: una experiencia de acompañamiento en la enseñanza de la Lengua escrita*[1]**,** el cual fue organizado por la Coordinación Estatal del Programa Nacional de Lectura, dependiente de la Dirección de Educación Básica de los Servicios de Educación Pública del Estado de Nayarit (SEPEN), en colaboración con investigadores de la Universidad Nacional Autónoma de México (UNAM), de la Universidad Autónoma Metropolitana Xochimilco (UAMX) y de la Escuela Nacional de Antropología e Historia (ENAH), bajo el financiamiento del Consejo Nacional de Ciencia y Tecnología (CONACYT).

El proyecto vincula diferentes especialidades que se apoyan entre sí para lograr un solo fin, el mejoramiento de la educación básica. Por tanto, aunque esta tesis sólo se enfocará en una parte del análisis lingüístico que se llevó a cabo, es importante describir lo que hizo la SEPEN para comprender mejor desde dónde parte nuestra tesis y cómo se relaciona con el proyecto.

Esta experiencia fue puesta en práctica en 100 grupos de primer grado y 80 grupos de segundo grado de los 7 Sectores Escolares de Educación Primaria del Estado de Nayarit. Participaron 36 asesores técnico-pedagógicos, 100 docentes frente a grupos de primer grado, 80 docentes frente a grupos de segundo grado y 4,200 alumnos.

En una primera fase se formó asesores técnicos- pedagógicos con conocimientos de didáctica del lenguaje para impulsar un ágil aprendizaje de la lengua escrita en los alumnos.

Posteriormente estos asesores se encargaron de capacitar a los 180 docentes que participaron, instruyéndoles sobre los procesos de alfabetización y los avances de investigación psicológica, lingüística, pedagógica y didáctica de la lengua. De esta forma a los asesores los consideraremos como *maestros acompañantes* y a los docentes que recibieron las capacitaciones los llamaremos *maestros acompañados*.

Después de cuatro años de trabajo se decidió hacer una evaluación del impacto de la formación de los maestros, así se realizó un ejercicio con los 4,200 alumnos que consistió en lo siguiente: cada maestro le leyó a su grupo la historia de una niña africana

[1] El proyecto está registrado en CONACYT con el número 50797 bajo la responsabilidad de la M. en C. Graciela Beatriz Quinteros Sciurano.

llamada Fátima. Esta historia consistía en una serie de imágenes y enunciados donde se describía la comunidad de la niña. Así, después de leerles el cuento se le pidió a cada uno de los niños que escribiera una pequeña narración sobre su vida, basándose en cuatro preguntas: ¿quién soy? ¿cómo es mi escuela? ¿cómo llego a mi escuela? y ¿qué hago en la escuela?

Finalmente, este último ejercicio también se llevó a cabo en escuelas donde los docentes no participaron en la experiencia[2]. Esto se hizo con el fin de recolectar todos los textos escritos por los niños y compararlos entre sí para ver de qué forma las capacitaciones de los maestros que sí participaron en la experiencia influyeron en el aprendizaje de los niños.

Como se puede suponer, de este trabajo se obtuvo una gran cantidad de textos, ya que aunque a unos niños les bastó una sola cuartilla para escribir, otros no pudieron usar menos de 6 para expresarse.

Para hacer el análisis lingüístico la SEPEN pidió la colaboración de la UAMX y del Instituto de Investigaciones Filológicas (IIFL) de la UNAM. Pero es de inferir que analizar una cantidad tan grande de textos de manera manual iba a hacer una tarea exhaustiva y prolongada, por lo que para poder analizar en un tiempo razonable el impacto que ha tenido la formación de los maestros, la UAMX y el IIFL solicitaron la colaboración del Grupo de Ingeniería Lingüística (GIL)[3], el cual es un equipo de investigación donde se tiene como objetivo el desarrollo de herramientas computacionales para realizar trabajos de búsqueda, reconocimiento, extracción de información, interpretación y reproducción del lenguaje humano[4].

Así, como integrante del GIL, el Mtro. Carlos Francisco Méndez Cruz diseñó, con ayuda de algunos análisis ya hechos por los lingüistas y supervisores[5], un programa computacional que facilitó el desarrollo del corpus electrónico y el análisis de los textos, así como también ayudó a obtener resultados más precisos que muestran de qué manera el proyecto *Aprender mientras se enseña: una experiencia de acompañamiento en la enseñanza de la Lengua escrita* impactó en el aprendizaje de los alumnos.

---

[2] A estos docentes los llamaremos *maestros no acompañados.*

[3] Dependiente del Instituto de Ingeniería de la UNAM

[4] El GIL es dirigido actualmente por el Dr. Gerardo Eugenio Sierra Martínez.

[5]Como responsable de este análisis lingüístico está la M. en C. Graciela Beatriz Quinteros Sciurano. Como supervisores tenemos a la Dra. Celia Zamudio de la UAMX, a la Dra. Celia Díaz del IIFL (UNAM), al Dr. Gerardo Sierra y al Dr. Alfonso Medina del GIL (UNAM). Los lingüistas y etiquetadores son María Inés Pucci y la Lic. Mercedes Tapia de la ENAH, la Lic. Argentina Robledo Domínguez de la UAMX y su servidora Sandra Nayely Richer Monroy del GIL (UNAM).

Como adición a esto es muy importante mencionar que el corpus que vamos a describir en esta tesis no es el primero que se desarrolla con bases computacionales en el GIL, como experiencia previa contamos con dos corpus más. El primero de ellos es el *Corpus Lingüístico en Ingeniería*, el cual fue dirigido por el Dr. Gerardo Sierra y el Dr. Alfonso Medina[6].

Este corpus es una recopilación de textos de las diversas áreas temáticas de la ingeniería y cuenta con una base bibliográfica que permite contabilizar las cantidades y tipos de documentos o palabras de manera automática. Para su elaboración participaron becarios y servidores sociales de la Facultad de Filosofía y Letras, de la Facultad de Ingeniería y de la Facultad de Contaduría y Administración.

Gracias al éxito de este primer corpus electrónico, y a los excelentes resultados obtenidos en un corpus no electrónico dirigido por la Dra. Concepción Company (IIFL)[7], se acordó elaborar un tercer corpus que implementa los recursos creados desde la filología y la ingeniería. De esta forma se desarrolló el Corpus Histórico del Español de México (CHEM)[8].

Como su nombre supone el CHEM es un corpus dedicado al análisis diacrónico, a diferencia de su predecesor, por lo que varias herramientas del Corpus Lingüístico en Ingeniería tuvieron que ser adaptadas a las nuevas necesidades. Pero independientemente de eso ambos corpus electrónicos, junto con el que describiremos en esta tesis, tienen muchas características en común, empezando por el lenguaje de programación con el que fueron creados.

De esta forma, aunque el proyecto *Aprender mientras se enseña: una experiencia de acompañamiento en la enseñanza de la Lengua escrita* sigue adelante con capacitaciones a los docentes, elaboración de nuevos materiales didácticos, actualizaciones del programa XML, etc. Esta tesis, que se enfoca en el análisis lingüístico desde su inicio hasta la actualización número 12ª del diseño del programa XML, describe un corpus confiable no sólo por las características que veremos más adelante, sino también porque lleva en sí experiencia y conocimientos adquiridos en

---

[6] El Corpus Lingüístico en Ingeniería está registrado como el proyecto *Desarrollo del Corpus Lingüístico en Ingeniería*, el cual fue patrocinado por CONACYT con el número R37712A bajo la responsabilidad del Dr. Gerardo Sierra. Asimismo este mismo corpus se encuentra registrado como el proyecto *Constitución de corpus lingüísticos electrónicos*, el cual fue patrocinado por DGAPA con el número IX402204 bajo la responsabilidad del Dr. Alfonso Medina.

[7] El corpus pertenece al proyecto *Generación de infraestructura filológica para la investigación y la docencia*, el cual fue patrocinado por CONACYT con el número 30873H.

[8] El proyecto del CHEM está registrado en DGAPA-UNAM, PAPIIT con el número IN400905 bajo el nombre *El Corpus Histórico del Español de México*.

otros corpus muy valiosos que han sido una aportación muy importante no sólo a la Academia, sino también a la evolución de métodos computacionales que de una u otra forma son de gran ayuda dentro de la sociedad en la que vivimos.

1.2. Planteamiento del problema

Debido a que esta tesis es parte del proyecto *Aprender mientras se enseña: una experiencia de acompañamiento en la enseñanza de la Lengua escrita*, nuestro problema y objetivos derivan por supuesto de algunos de los del proyecto.

Por tanto, el problema de nuestra tesis es la necesidad de contar con un corpus etiquetado de discurso infantil escrito que permita realizar distintos tipos de análisis; por ejemplo, a nivel de palabra se requiere estudiar el uso de mayúsculas, el uso de signos de puntuación, la manera de segmentar las palabras, la forma en cómo se autocorrigen, el uso de las alternancias gráficas como “g”, “j”, “c”, “s” y “z”, entre otros fenómenos.

Este problema surge debido a que en la actualidad se han realizado pocas investigaciones relacionadas con la adquisición del lenguaje escrito y menos aún sobre el español de México.

Así, nuestro problema en esta tesis es contar con una herramienta que nos proporcione información actual sobre los fenómenos ortográficos más frecuentes en la adquisición de la lengua escrita de niños mexicanos.

1.3. Objetivo general

El objetivo de la tesis es describir una metodología para elaborar un corpus etiquetado de discurso infantil escrito que muestre en cuáles y en cuántas palabras se observan fenómenos ortográficos. Indicando por supuesto el tipo de éstas (omisiones de grafías o acentos, sustituciones, agregados o inversión en el orden de las grafías, etc.).

1.4. Objetivos particulares

a) Describir los objetos lingüísticos que se analizaron en el proyecto *Aprender mientras se enseña: una experiencia de acompañamiento en la enseñanza de la Lengua escrita*.

b) Estudiar lo que es el discurso escrito, el discurso oral y las diferencias que hay entre ellos. Esto con el propósito de vislumbrar la importancia de analizar corpus como el que describiremos en esta tesis.
c) Estudiar otras investigaciones que trabajan con la lingüística computacional para estar al tanto del uso de otras herramientas computacionales en la lingüística, de las metodologías que se han desarrollado para otros fines lingüísticos y de los resultados que se han logrado obtener. Esto se hará con el objeto de tener un panorama más amplio de lo que es la lingüística computacional y de considerar algunas ideas que puedan aportar información importante para los resultados y conclusiones de esta tesis.
d) Estudiar lo que es un corpus y describir la elaboración del corpus del proyecto.
e) Analizar los resultados del objeto de estudio y obtener los porcentajes que ocupan los fenómenos en éste.

1.5. Objeto de estudio

Como hemos mencionado, en el proyecto se hizo una recolección de textos escritos por niños, algunos pertenecientes a escuelas dónde los maestros fueron capacitados, y otros pertenecientes a escuelas donde los maestros no fueron capacitados. De esta recolección se tomaron alrededor de 300 textos[9] escritos por niños de segundo año de primaria y se repartieron entre los lingüistas para ser analizados[10]. Así, el objeto de estudio de esta tesis son 121 textos, es decir, los textos que me fueron asignados para trabajar, de los cuales 60 de ellos fueron escritos por alumnos de maestros acompañados y 61 textos fueron escritos por alumnos de maestros no acompañados.

1.6. Estructura de la tesis

Esta tesis se desarrolla en seis capítulos. Este primero consiste en una introducción donde se da una breve explicación sobre el proyecto en el que se enmarca esta tesis, mencionando por supuesto a las instituciones y las personas relacionadas, asimismo, se explica cuál es el problema de esta tesis, se desarrollan sus objetivos a lograr, se

[9] La mitad de los textos pertenecen a niños de maestros que participaron en la experiencia, y la otra mitad a niños cuyos maestros no participaron.
[10] De todo el acervo de textos que se recolectó, estos primeros 300 textos que se analizaron constituyen el corpus del proyecto.

describe el corpus en el que se basó el proyecto y el objeto de estudio en el que se basa esta tesis.

En el segundo capítulo se verá lo que es el discurso, las características particulares del discurso escrito y se estudiará la forma en que los niños adquieren el discurso escrito de la lengua materna. Asimismo, se describirán otras investigaciones que se relacionan con el proyecto *Aprender mientras se enseña: una experiencia de acompañamiento en la enseñanza de la Lengua escrita*, y que analizan algún aspecto del discurso infantil utilizando herramientas computacionales. Por supuesto que los aspectos que analizan y las herramientas de las que se sirven no son los mismos que en este proyecto, pero coincidimos en aprovechar lo que la lingüística computacional nos ofrece para estudiar y proponer cosas nuevas, que no sólo benefician a la investigación académica, sino que también pueden beneficiar e impulsar el desarrollo social en varios aspectos.

En el tercer capítulo se analizará qué es un corpus y cómo se constituye, se explicarán las diferencias entre corpus manuales y electrónicos, también veremos lo que es anotación y los tipos que hay, y finalmente se hablará en breve sobre lo que es el XML. Este capítulo es muy importante porque nos ayudará a entender lo que es una etiqueta y cómo se codifica su significado, y así podremos comprender mucho de lo que vendrá en los siguientes capítulos.

Posteriormente, en el capítulo cuatro se desarrollará lo que es el pilar principal de esta tesis: la construcción del corpus; en ella se describirán los pasos que se siguieron para constituir el corpus, se explicará qué es el preprocesador de archivos para etiquetado XML, y de qué forma fue útil. Asimismo, se especificará el significado y función de cada una de las etiquetas que se utilizaron, y finalmente explicaremos cómo nos cercioramos de que el trabajo manual se hizo bien.

En el quinto capítulo se expondrán los resultados obtenidos: se describirá lo que es la hoja de estilo, analizaremos el porcentaje que ocupan los fenómenos ortográficos en el objeto de estudio de nuestra tesis y en base a los conocimientos adquiridos en la universidad se tratará de comprender y describir posibles razones que nos expliquen la existencia de algunos fenómenos encontrados.

En el capítulo seis se harán algunas sugerencias para mejorar el esquema del corpus y se propondrán posibles trabajos a futuro que se relacionan con este trabajo.

Finalmente se presentará la bibliografía que se utilizó.

2. DISCURO INFANTIL ESCRITO

Como hemos mencionado, el corpus está basado en textos escritos por niños, pero antes de pasar al análisis es importante tener en cuenta algunos conceptos para entender mejor sobre qué bases teóricas se trabajó; por tanto, en este capítulo veremos lo que es el discurso, después nos centraremos en las diferencias que tiene el discurso escrito con respecto al discurso oral. Posteriormente analizaremos un aspecto más concreto: la adquisición de la escritura de los niños y finalmente describiremos algunas investigaciones interesadas también en este último aspecto.

2.1. Discurso

A lo largo de la historia han existido varias escuelas que se han preocupado por el discurso, y aunque éstas coincidan en buscar la mejor definición y técnica de análisis, sus teorías disciernen en varios aspectos debido a los diversos puntos de vista: sociológico, psicológico, lingüístico, comunicativo, etc.

De acuerdo con Pardo el discurso es "una unidad lingüística en la que un mensaje es expresado en un enunciado que supera la oración" (2007, pág. 38), es decir, el discurso es una expresión basada en un enunciado compuesto por una o varias oraciones, pero este enunciado va más allá de la oración debido a que, de acuerdo con Benveniste, tiene una intensión comunicativa (Maingueneau: 1976, pág. 10). Asimismo, dependiendo del lugar donde se lleve a cabo, el discurso seguirá las reglas correspondientes de una lengua determinada, logrando a la vez ser un reflejo de las *tradiciones discursivas* (entendiendo a estas últimas como las características de comunicación de una región en particular, las cuales se llevan a cabo gracias a los conocimientos socioculturales que comparten los habitantes). De esta forma se puede considerar como discurso la narración de un libro, una noticia, una entrevista, una clase, una carta, etc.

La *variación comunicativa* es un fenómeno inherente al discurso en razón de su pluralidad, interacción situacional y a la variedad de condiciones discursivas (Oesterreicher: 1996). Como parte del acto comunicativo el emisor está en una constante formulación y verbalización de las ideas, pero para esto, además de tener en claro las reglas del lenguaje que está utilizando y las tradiciones discursivas, su discurso

depende de todos los elementos de la variación comunicativa, es decir, las palabras que formula y la forma en cómo las expresa dependen de la situación y el contexto en el que se encuentre, así como de su interpretación de la realidad extralingüística.

Pero a pesar de esta variación comunicativa Wulf Oesterreicher (1996) enumera algunos parámetros universales que ayudan a encasillar en un nivel extralingüístico los diferentes discursos que se puedan dar, así como también ayudan a caracterizar las tradiciones discursivas:

1. Todo discurso tiene un grado de *privacidad*, el cual depende de la cantidad de participantes que actúen, de esta forma entre menos participantes haya el discurso será más privado.
2. El discurso puede desarrollarse de distinta manera dependiendo del grado de *intimidad* entre los participantes, es decir, qué tanto se conocen entre ellos y cuánta confianza hay.
3. Así, dependiendo de la privacidad y la intimidad tendremos un grado de *emotividad*, siendo esta menor, por ejemplo, si se trata de un discurso científico, o mayor si se trata de un discurso entre amigos.
4. Dependiendo del contexto, el discurso puede ser muy corto y conciso o muy extenso y detallado, así, en el primer caso se podría considerar un grado bajo de *inserción* o *implantación*, y en el segundo caso habría un grado alto.
5. Los discursos siempre hacen *referencia* a algo, pero el grado de ésta dependerá de que tan próxima o lejana se encuentra la persona o el objeto al que se hace mención en el discurso.
6. La *distancia temporal* o *local* que haya entre los locutores define el grado de contacto. Por ejemplo, si yo leo *Don Quijote de la Mancha*, la distancia temporal y local entre Cervantes y yo tendrá un bajísimo grado de contacto.
7. Dentro de un discurso, la cantidad de expresiones orales o corporales del receptor definen el grado de *cooperación*.
8. El grado de *dialoguicidad* se define por las veces en que se alternan entre los participantes el papel de locutor.
9. De acuerdo con prácticamente todo lo anterior se puede definir también el grado de *espontaneidad* que haya entre los participantes.

10. Finalmente, dependiendo si el discurso gira en torno a un solo tema o a varios se podrá considerar un grado de *fijación* o *determinación*. Así, si la conversación gira en torno a un solo tema, el discurso tendrá un alto grado de fijación.

Con todos los puntos anteriores no estamos diciendo que cada uno de los discursos se caracteriza por uno solo de estos parámetros, más bien la clasificación (o perfil comunicativo) de los discursos se define por la cantidad y combinación de parámetros que los caractericen así como de la delimitación del contexto, el cual influirá de manera considerable en la aparición y grado de éstos. De esta forma Wulf nos define cuatro tipos de contexto:

1. El *contexto situación* se caracteriza porque los interlocutores y el referente se encuentran dentro de la misma situación comunicativa, es decir, en el mismo espacio y tiempo.

   Es importante aclarar que al hablar de **espacio** no nos estamos refiriendo únicamente a un lugar determinado donde se encuentran dos o más personas, sino también a espacios cibernéticos, telefónicos o masivos que la tecnología ha abierto por medio del internet, celulares, teléfonos, radios, etc. En estos casos, aunque las personas no se encuentren en un mismo lugar, comparten un mismo espacio.

   Asimismo, al hablar de **tiempo** nos referimos al momento en que el mensaje es reproducido y recibido, entonces cuando los participantes están en un mismo tiempo, es decir, en la misma situación comunicativa, en cuanto el mensaje es emitido es recibido por el receptor.

2. El *contexto cognoscitivo* se caracteriza por dos aspectos: por los conocimientos individuales que tiene un interlocutor y los conocimientos que comparte con los demás; y por el conocimiento que tienen los interlocutores sobre información universal y sobre las tradiciones discursivas de la lengua que están utilizando.
3. El *contexto lingüístico de la enunciación* se constituye por expresiones lingüísticas que incluyen, preceden o siguen una frase verbal. Estas expresiones contribuyen para desambiguar el significado de la frase.
4. El *contexto comunicativo* puede dividirse en dos: *paralingüístico* y *no-lingüístico*. El primero se caracteriza por los fenómenos prosódicos que se

presenten en el discurso; y el segundo se refiere a las condiciones y circunstancias de la interacción comunicativa.

Es importante mencionar que como parte de un discurso inmediato se pueden presentar prácticamente todos los contextos que se mencionaron, en cambio, cuando la comunicación no es inmediata el contexto sólo puede ser enunciativo y/o cognoscitivo[11]. Por tanto, gracias a la variedad de contextos, el discurso inmediato se caracteriza también por una economía de lenguaje, y si añadimos un alto grado de espontaneidad e intimidad, podremos encontrar en este discurso una gran cantidad de vacilaciones, repeticiones, saltos, etc. En cambio, debido a la poca cantidad de contextos participantes, el discurso no inmediato se caracteriza por ser más explicito y detallado con la información que proporciona, por lo que generalmente suele ser planificado y en consecuencia carece de vacilaciones, repeticiones y demás elementos.

2.2. Discurso escrito

Cabe considerar que la oralidad antecede a la lengua escrita, debido a que todas las comunidades desarrollan el discurso oral y algunas posteriormente desarrollan el discurso escrito. Es por ello que dentro del análisis lingüístico anteriormente se consideraba al discurso oral en una posición superior al discurso escrito. Fue hasta apenas en el siglo XX (Albeentosa: 2001) cuando, gracias a los nuevos recursos tecnológicos se llevaron a cabo trabajos sobre cada discurso y con éstos se concluyó que ambos discursos tienen el mismo valor aunque hay características sobresalientes que definen a cada uno de ellos como único e independiente.

La primera característica es por supuesto la forma física: el discurso escrito está formado por grafías y el discurso oral por fonemas, y aunque es obvia esta observación, realmente es muy importante, ya que nos permite visualizar que entonces ambos discursos son sistemas que se relacionan pero que son independientes, por tanto, “la escritura (…) se organiza según reglas propias que obedecen a circunstancias

[11] Wulf considera que el discurso inmediato es el discurso oral y el discurso no inmediato se lleva a cabo por medio de la escritura, pero si tomamos en cuenta las conversaciones que se dan por el chat, la escritura puede desarrollar un discurso inmediato, por lo que entonces podemos discernir de esta clasificación un poco radical para la época en la que vivimos. Asimismo, Wulf considera que la distancia entre los interlocutores puede excluir algunos contextos, pero si retomamos la experiencia del chat, sabemos que con ayuda de audífonos, cámara y micrófono en un discurso a distancia se pueden presentar todos los contextos. Por tanto, como se puede observar, sólo menciono el discurso inmediato y no inmediato, pero procuro no clasificar ninguno de ellos como discurso escrito u oral.

específicas de producción[12] y transmisión diferentes a las del discurso oral" (Pérez: 1995, pág. 62).

Otra característica es que la escritura da una imagen de homogeneidad a la lengua, por ejemplo, un texto escrito en español no guarda en sí las distintas pronunciaciones que tienen diferentes hablantes como argentinos, españoles, mexicanos, peruanos, etc. Asimismo, los textos antiguos también guardan una imagen históricamente homogénea.

Con esto último tocamos otra característica muy importante: la escritura tiene un carácter de almacenamiento, y gracias a esto se han conservado algunos viejos escritos que son la única evidencia de lenguas antiguas. En la antigüedad no había forma de hacer registros orales, por lo que la escritura ha sido la única herramienta que han tenido los estudiosos para analizar y comprender aquellas lenguas que actualmente ya no existen o se usan muy poco como el latín. Asimismo, algunos escritos han sido las únicas huellas sobre las cuales se ha logrado comprender y seguir la evolución por la que han pasado las lenguas actuales.

Pero el carácter de almacenamiento no sólo sirve para poder conocer las lenguas antiguas y seguir su evolución, sino que en la actualidad ha servido también "para establecer el conjunto de las normas que rigen los usos de las lenguas" (Blanche: 2002, pág. 27). Asimismo, este establecimiento de normas forma parte de la homogeneidad de la escritura, ya que vincula épocas y regiones alejadas. Por ejemplo, el almacenamiento ayuda a que haya una homogeneidad del español, la cual vincula el español de México con el de España y otros países, así como también vincula el español del México actual con el del México antiguo.

Asimismo, además de la vinculación temporal y regional, tener normas ortográficas ayuda muchas veces a eliminar algunas ambigüedades que se pueden presentar en el discurso escrito, por ejemplo, nos puede ayudar a definir la frase /aser/ como *hacer* o *a ser*, o definir la función de /a/ como verbo (ha) o preposición (a), etc.

Con todo esto es importante mencionar que el almacenamiento no pretende mantener un anclaje total de la escritura, ya que ésta, al igual que el discurso oral, también evoluciona para adaptarse a los nuevos sistemas. Y aunque el discurso escrito formal muchas veces no acepta ciertas palabras y frases coloquiales, éstas pueden expresarse en una comunicación informal, a tal grado que incluso la ortografía y algunas

[12] Donde se requiere el uso de herramientas como papel, lápiz, pluma, máquina de escribir, computadora, celular, etc. (Harris, Roy. 1999. Pág. 46).

reglas normativas de la escritura son infringidas. Así, por ejemplo, es común encontrar en conversaciones cibernéticas, en redes sociales, o mensajes de celular frases como *pus* (pues), *pa'* (para qué), *on tas* (dónde estás), tmb (también), x' k (porque o por qué), tqm (te quiero mucho), etc. Pero esto no sólo se da en contextos cotidianos, sino también en contextos literarios, ya que hay autores como Vallejo, Huidobro, Cortázar, García Márquez, entre muchos otros, que no sólo usan coloquialismos, sino que además inventan palabras o juegan con la sintaxis u otros aspectos gramaticales para darle otro valor al arte literario.

Además, continuando con la informalidad del discurso, así como hay elementos en la oralidad que la escritura no puede transcribir, también hay elementos en la escritura que no se pueden pronunciar, por ejemplo algunas descripciones de emociones como :) (alegría), :D (mucha alegría), :P (sarcasmo, broma, juguetón, …), :X (error, avergonzado, tonto, …), etc.

Por supuesto en el discurso escrito formal también podemos encontrar elementos que no se pueden pronunciar como los números de las notas, los asteriscos, paréntesis, comillas, etc.

De esta forma es como el carácter de almacenamiento de la escritura aporta muchas ventajas sin convertirse en un impedimento para que el discurso escrito sea flexible, además, fuera de los estudios lingüísticos, este carácter de la escritura nos ayuda también con actividades cotidianas como hacer contratos, conservar cuentos, historias o noticias para releer una y otra vez, conservar composiciones que pueden reproducir varias personas, conservar apuntes que se pueden estudiar y repasar varias veces, etc.

Finalmente, en relación con este complejo carácter, hay que mencionar que actualmente podemos considerarlo como opcional en la escritura, debido a que en ciertos contextos su existencia depende de la decisión del escritor. Para explicar mejor esto pongamos como ejemplo las conversaciones escritas por medio del Internet: su carácter de almacenamiento dura mientras la ventana de conversación está abierta, pero una vez que ésta se cierra todo lo que se escribió desaparece y no se puede recuperar a menos que uno de los participantes haya decidido guardar la conversación. Asimismo, todo lo que se escriba en archivo electrónico (Word, TXT, Excel, etc.) se conservará sólo si el escritor lo desea.

Otra característica que también puede resultar algo obvia, es que la escritura es visual, lo cual llega a hacer muy útil para hacer algunos análisis o cálculos que no se

pueden hacer en la oralidad, por ejemplo cálculos matemáticos, físicos, químicos, etc. Pero también, desde un punto de vista lingüístico, el poder visualizar a la lengua ha permitido llevar a cabo análisis a nivel léxico, morfológico, sintáctico, semántico, etc., debido a que el discurso oral sólo había permitido llevar a cabo análisis fonéticos y fonológicos (Albeentosa: 2001). De esta manera, la forma física del discurso escrito ha permitido tener un conocimiento más detallado de la lengua, ya que permite estudiarla y analizarla detenidamente.

Además de todo lo anterior, el discurso escrito se destaca por muchas otras características, algunas de las cuales son resumidas por Pérez Grajales (1995, pág. 62) en la siguiente lista:

1. Sustituye la deixis implícita en la enunciación oral por la descripción detallada del emisor-receptor: rol, gestos, movimiento del cuerpo, las manos y la cara.
2. Utiliza la acentuación y la puntuación para sustituir las pausas y la entonación es decir, lo que se denomina código paralingüístico.
3. La comunicación escrita está limitada a situaciones reducidas de la vida (alfabetizados), mientras la comunicación oral es universal.
4. La escritura elude la redundancia léxica, sintáctica y semántica como recurso de eficacia y trata de evitarla al máximo como recurso de comprensión. De ahí que se imponga el uso de sinónimos o la elipsis como medios anafóricos de cohesión.
5. La escritura exige una planeación cuidadosa que involucre la macroestructura, la superestructura y la formulación de enunciados lingüísticos que orientarán la composición del texto completo. Tradicionalmente, este aspecto se ha dejado al azar y de ahí las fallas que afectan la coherencia global.
6. Hay exigencias de acatamiento de las reglas sintácticas y semánticas. Se censura socialmente su violación y se exige un uso óptico de la lengua de acuerdo con la norma estándar. Es por esto que siempre se han corregido las fallas ortográficas, de puntuación, las discordancias entre sustantivo y adjetivo, (…).

Aunque considero que estas características son muy acertadas me gustaría hacer dos aclaraciones:

- Con respecto al punto número uno, las historietas pueden utilizar escritura de "segundo grado", de esta forma no sólo los personajes ayudan a que haya una expresión corporal, sino que además podemos leer risas, golpes, estornudos, tos,

etc. Aunque por supuesto aún quedan excluidos elementos como prolongaciones, aspiraciones, alturas de voz, etc. (Blanche: 2002, pág. 16).

- Si retomamos las conversaciones del internet y los celulares, entre otros, veremos que lo que dicen los puntos cinco y seis no aplica.

2.3. Adquisición del conocimiento del lenguaje escrito

Al entrar en temas un poco más especializados como este, hay que aclarar que sólo se describirá la adquisición del discurso escrito de niños con capacidades físicas y mentales que se encuentran dentro de parámetros normales o estándares, ya que la adquisición del discurso se realiza de distinta forma y bajo diversas circunstancias en niños con alguna incapacidad física o mental. Asimismo, sólo analizaremos la adquisición del lenguaje escrito en niños monolingües.

De acuerdo con Ferreiro y varios investigadores más, el nivel de conceptualización de la escritura que tengan los niños desde antes de aprender a leer y escribir, y la clase social a la que pertenezcan influirá de una u otra forma en la adquisición de la escritura, debido a que estos factores pueden convertirse en un impulso o un obstáculo para que el niño fácilmente adquiera el lenguaje escrito. (Ferreiro: 1982)

Con *nivel de conceptualización* nos referimos al conocimiento que tienen los niños sobre la escritura, el cual es construido por ellos mismos dependiendo del ambiente donde se desarrollen, y de las herramientas o ayuda que se les proporcione.

En la medida en la que los niños se van relacionando con el medio que los rodea construyen hipótesis de lo que ven, tocan, escuchan, etc., para tratar de comprender lo que ocurre a su alrededor, así, de la misma manera, al crecer en una sociedad donde el discurso escrito tiene una presencia constante, los niños se van formulando hipótesis sobre éste antes de que alguien les enseñe a leer y escribir, y de acuerdo con qué tan cercanos estén los niños de la escritura podrán avanzar con mayor o menor velocidad en sus niveles de conceptualización. Generalmente[13] las zonas urbanas son favorables para que los niños generen esta conceptualización a temprana edad, ya que son las zonas donde el discurso escrito está presente de muchas formas: señalamientos, letreros de publicidad, volantes, carteleras, instructivos, etc. En cambio, en las zonas rurales la

[13] En esta explicación estamos generalizando conscientes de que siempre hay excepciones.

presencia del discurso escrito es menor, por lo que el nivel de conceptualización que tengan los niños de estas zonas puede ser bajo o nulo antes de que entren a la escuela.

Asimismo, independientemente de la zona, la vinculación que tengan los padres (o tutores) con la escritura también influye en los niños, ya que si los padres tienen el hábito de lectura o la escritura es una constante en actividades cotidianas, el niño estará relacionado con un ambiente que le ayudará también a generar una idea de lo que es el discurso escrito. Estos tipos de ambientes familiares generalmente se desenvuelven en una *clase social* media o alta, donde uno o ambos padres tienen estudios superiores a la secundaria. En cambio, dentro de la clase baja el discurso escrito no es muy presente en los hogares, ya que los límites económicos obligan a las personas a adquirir alimento, vestimenta, etc., antes que un libro, una revista, juegos con letras, y más. Así también, generalmente en esta última clase social los padres no tienen estudios superiores a la secundaria, por lo que su vinculación con la escritura y su interés por la lectura es muy baja.

De esta forma, si el niño está rodeado de uno o varios factores positivos construirá una idea de lo que es la escritura y cuál es su posible uso, y gracias a esto, cuando llegue a la escuela, aprenderá a escribir y leer de forma rápida y sencilla. En cambio, un niño que por alguna razón no tuvo la motivación para conceptualizar la escritura desde temprana edad, al estar frente a ella tendrá algunas dificultades y tardará un poco para adquirirla.

Para cualquiera de los dos casos, una vez que los niños entran a la escuela, lo que favorecerá una adquisición de la escritura exitosa es una buena metodología de enseñanza, no sólo porque ésta ayudará al niño a codificar lo que lee, sino porque además le ayudará a comprender el texto, lo cual es la finalidad primordial.

Desde muy pequeños los niños comienzan a hacer garabatos, en una primera etapa el niño sólo se interesa por las líneas que hace, pero ya en una segunda etapa estas líneas comienzan a tener un significado para el niño, por lo que comienza a atribuirles nombres. De esta forma es como los niños comienzan a hacerse hipótesis sobre lo que es la escritura, y los garabatos son las primeras grafías que expresan esa construcción de conocimiento (Sinclair: 1982, pág. 96).

Posteriormente de que los garabatos adquieren significado, el niño trata de imitar la escritura haciendo garabatos ligados en una cadena para imitar la letra cursiva, o garabatos separados para imitar la letra de imprenta. Asimismo, en esta etapa la producción de los dibujos y de la escritura se confunde bastante, debido a que el niño

aún no delimita las diferencias entre las líneas que componen un dibujo y las que construyen letras, incluso muchas veces los garabatos- letras parecen ser accesorios o partes del dibujo.

El texto representado por los garabatos- letras no "dice" nada y en consecuencia no tiene significado alguno, éste se adquiere únicamente gracias a algún dibujo debido a que en esta etapa los garabatos- letras sólo "guardan una relación de pertenencia tan frágil que ella se desvanecería si la inclusión dentro de los límites de la figura no la garantizara. Paulatinamente la escritura, para no confundirse con el dibujo, tiende a salirse fuera de los límites de éste" (Ferreiro: 1982, pág. 132), pero el significado del texto sigue dependiendo del dibujo.

Así, por el hecho de vincular siempre el texto con una imagen, muchas veces los niños escriben los garabatos- letras de tal forma que en su forma física guardan alguna relación con la imagen o el referente en cuestión. Por ejemplo, en un ejercicio que describe Downing (1982, pág. 242), a un niño se le pidió que escribiera *pato* y éste hizo una serie de garabatos, después se le pidió al mismo niño que escribiera *oso* y éste hizo la misma serie de garabatos sólo que más grandes, cuando se le preguntó al niño por qué había hecho garabatos pequeños para la palabra *pato* y garabatos grandes para la palabra *oso*, el niño respondió que porque un oso es más grande que un pato. De esta forma toda representación produce una imagen y por tanto, el garabato es la imagen de un primer proceso de comprensión.

Con el ejemplo anterior hemos visto además otra cosa muy importante que ocurre en varios casos: lo que escriben los niños con sus garabatos- letras sólo puede ser comprensible por ellos mismos, y ellos lo saben, por tanto, como parte de la construcción de lo que suponen que es la escritura, los niños deducen que la escritura sólo puede ser legible y comprensible para quien lo escribe, es decir, "cada uno puede interpretar su propia escritura pero no la de los otros" (Downing: 1982, pág. 242).

En una siguiente etapa, con base en la conceptualización de escritura que los niños se han formulado, ellos ya pueden deducir qué se puede leer y qué no. De esta forma, según un estudio realizado por Ferreiro se concluyó que en esta etapa los niños consideran que para que una palabra pueda ser leída tiene que tener por lo menos tres letras o garabatos- letras, es decir, para ellos palabras de una o dos letras como las preposiciones o artículos no pueden ser leídas y en consecuencia no tienen significado alguno.

Asimismo, en esta etapa ocurre algo muy importante: los niños se dan cuenta que una palabra necesita variación de grafías. En este momento los niños dejan de hacer cadenas de garabatos y comienzan a crear pequeños códigos de garabatos- letras o a utilizar por lo menos las letras de las vocales y/o las de sus nombres para escribir[14].

En relación con el texto e imagen, en esta etapa el texto ya "dice" algo, es decir que ya tiene significado y puede ser tratado por el niño como sistema independiente del dibujo, pero su significado sólo puede ser relativo a una imagen próxima o a algún referente, en otras palabras, el texto ya no es sólo un símbolo de pertenencia, ya guarda en sí un significado, pero éste sigue relacionándose con algún dibujo. Asimismo, dentro de una palabra cada una de las letras significa lo que significa toda la palabra, pero cuando se encuentran aisladas no tienen valor alguno, porque para los niños los garabatos- letras o las letras no "dicen" nada "o, en todo caso, sólo pueden decir lo que ellas mismas son: letras" (Ferreiro: 1982, pág. 137).

Posteriormente en otra etapa, el niño se da cuenta que las letras tienen una relación sonora con el habla, por tanto, comienza a atribuirle a cada letra (o garabato-letra) una sílaba, lo cual no debe de sorprendernos, ya que como reafirma Sinclair (1982), las sílabas son las unidades básicas del lenguaje oral. De esta forma, en esta etapa cada una de las letras ya no conservan el valor total de la palabra, sino que comienzan a tener un valor y un significado por sí mismas. Así, por ejemplo, la grafía "I" puede valer por el conjunto de fonemas /ka/, y la grafía "E" puede valer por el conjunto /sa/, logrando así que IE signifique *casa*.

A pesar del gran avance que ha logrado el niño, tarde o temprano comenzará a tener problemas debido a que las palabras mono o bisilábicas deberán estar compuestas por una o dos letras correspondientemente, y esto chocará con su teoría de que una palabra requiere mínimo de tres letras. De esta forma, en una siguiente etapa, el niño requerirá ir más allá de las sílabas y es en este punto donde hay que presentarle el alfabeto como la solución a su problema.

Al aprender el alfabeto, el niño tiene que aprender qué sonido le corresponde a cada grafía, pero como señala Luis Fernando Lara (2002, pág. 53), las grafías que usamos representan sonidos secundarios, no representan el extenso repertorio de fonemas que pronunciamos día a día y que los niños perciben fácilmente, por lo que no debería sorprendernos que a veces la escritura de los niños parezca una transcripción

[14] Aunque usen letras del alfabeto no significa que escriban de acuerdo con la norma, en esta etapa ellos puede escribir algo como "iae" para decir *carro*.

fonética y encontremos fenómenos como “miscuela”, “enlatard”, “ballos”, “las niñason”, etc. Esto se debe a que auditivamente para el niño hay líneas muy delgadas entre ciertos fonemas como *ll* y *ñ*, *r* y *l*, o *p* y *b*, etc., por lo que le cuesta trabajo encasillar cada sonido en una letra y además separar las palabras cuando en la oralidad nunca lo hacemos.

Finalmente en una etapa más evolucionada ocurre otro fenómeno muy curioso:

> (…) los niños son muy selectivos respecto de lo que se puede leer y escribir. (…), en primer lugar creen que sólo deben escribirse las palabras que nombran cosas o personas, más adelante aceptarán también las que representan acciones, pero todavía se omiten los artículos y preposiciones, que pasarán a escribirse en forma independiente mucho más tarde. (…) la actividad del dictado, realizado por Ferreiro (1982) muestra de qué modo los niños aprenden también en la escuela que no todo lo que el maestro dice cuando dicta se debe escribir: hay que omitir los comentarios y recomendaciones. (…) los niños creen que se puede decir pero no escribir cuando uno se refiere a algo que no existe, que es falso y que es imposible (Pontecorvo: 2002, pág. 135).

De esta forma terminamos este apartado, exponiendo la adquisición de la escritura en sus primeras etapas sabiendo que ésta se sigue perfeccionando conforme los niños la utilizan, es decir que a partir de lo que hemos descrito, los niños van modificando cada vez más su conceptualización de escritura.

2.4. Estudios sobre discurso infantil escrito

Como ya hemos mencionado antes existen pocos estudios sobre la adquisición de la lengua escrita en español, y dentro de éstos podemos encontrar algunos que se han servido de herramientas computacionales al igual que la investigación en la que se basa esta tesis. Así, en este punto expondremos dos investigaciones que se sirven de la tecnología computacional para analizar algunos aspectos del discurso infantil escrito. Esto lo hacemos con el propósito de conocer otras metodologías de estudio y las diferentes herramientas que pueden auxiliarnos para analizar en español.

Además, al ser investigaciones semejantes al de esta tesis, pretendemos que sus resultados y conclusiones nos ayuden a darle explicación a algunos de nuestros resultados cuando lleguemos a ellos.

Asimismo, en este punto describiremos un sistema que ha recolectado una gran cantidad de textos escritos, orales y audiovisuales en varios idiomas, y los ofrece al público junto con algunas herramientas computacionales para que quien lo desee pueda llevar a cabo algunos análisis lingüísticos con ese material.

2.4.1. Caperucita Roja aprende a escribir

Caperucita Roja aprende a escribir es el título de un libro en el cual se describe un proyecto llamado *La adquisición de la lengua escrita en diversos contextos lingüísticos y educativos*. Este proyecto está dedicado a la investigación de la adquisición de la lengua escrita y fue organizado por la Dra. Clotilde Pontecorvo y la Dra. Emilia Ferreiro, con ayuda del patrocinio bilateral del consejo de investigación científica de México (CONACYT) y de Italia (CNR).

La razón por la que se realizó este proyecto fue debido a que hay pocos estudios sobre cómo los niños aprenden a escribir su lengua materna. Además que la mayoría de esos estudios tienen bases puramente teóricas y no hacen un análisis desde una perspectiva psicolingüística y cognitiva como se está haciendo actualmente en muchos estudios sobre la oralidad. Por tanto, en este proyecto se decidió hacer un estudio sobre la adquisición de la lengua escrita desde esa perspectiva psicolingüística y cognitiva debido a que ésta permite analizar el discurso escrito tanto en su desenvolvimiento dentro de la escuela como fuera de ella.

Gracias a los sorprendentes resultados de un ejercicio comparativo[15] que se llevó a cabo antes de este proyecto, se convino recolectar textos escritos por niños mexicanos, uruguayos, brasileños e italianos, debido a que la comparación entre las ortografías de cada uno de los idiomas de estos niños permitiría vislumbrar muchos fenómenos que no se encontrarían al analizar una sola lengua.

---

[15] El ejercicio consistió en "confrontar niños italianos e hispanos para verificar el grado de generalidad de (…) (la) «exigencia de variedad interna»" (Ferreiro: 1996, pág. 18).

Como hemos dicho, antes de que los niños aprendan a leer y a escribir ya tienen una idea de cómo pueden o deben estar constituidas las palabras, así, en el libro Los sistemas de escritura en el desarrollo del niño (Ferreiro: 1982) se describe un ejercicio que consistió en que a los niños se les daba conjuntos de palabras como "AAMM" y "AMEM", y se les preguntaba cuál podría ser una palabra. En general los niños escogían el segundo ejemplo porque consideraban que una palabra debía tener variedad de letras, es decir, *variedad interna*.

Antes de llevar a cabo el proyecto *La adquisición de la lengua escrita en diversos contextos lingüísticos y educativos*, se hizo el ejercicio que acabamos de describir con niños italianos e hispanos, y a pesar de que en el italiano son comunes las palabras con duplicidad de letras, los resultados concluyeron que los niños italianos son quienes exigen más la variedad interna.

Antes de empezar con la recolección del corpus se decidieron los siguientes tres puntos: Primero, lo más conveniente era crear una base de datos "que permitiera una comparación sistemática, sobre parámetros diversos, de textos de niños de diversas edades, desde el comienzo de la escritura de tipo alfabético" (Ferreiro: 1996, p. 20). Segundo, se acordó que los niños escribieran la historia de Caperucita Roja por dos razones: a) Es una historia conocida en México, Uruguay, Brasil e Italia. b) "facilita la tarea de los escritores debutantes, ya que no deben asumir la carga cognitiva de inventar una historia razonable" (Ferreiro: 1996, p. 20). Y tercero, se convino utilizar una herramienta computacional que ayude a identificar varios problemas interesantes. Para esto, se solicitó la ayuda de Isabel García Hidalgo quien desarrolló un sistema para hacer análisis a nivel lingüístico, este sistema actualmente es conocido en México con el nombre de TEXTUS.

Una vez que se consiguieron los textos escritos por los niños, se construyó el corpus electrónico para ser analizado con ayuda del TEXTUS, logrando así obtener los siguientes resultados:

1. Dentro de la segmentación existen dos fenómenos: la hiper y la hiposegmentación. La primera consiste en que los niños separan una palabra en dos o más segmentos y la hiposegmentación consiste en que los niños pegan dos o más palabras como si fuesen una sola. Con respecto a estos fenómenos, el análisis reveló que los niños italianos son los menos propensos a cometer estos errores, seguidos de los niños uruguayos, después los brasileños y finalmente los mexicanos.
2. Las tres lenguas (español, italiano y portugués) tienen en común que en cada una de ellas la hiposegmentación es más frecuente que la hipersegmentación.

Una de las explicaciones que se deriva entonces de estos dos puntos es que posiblemente los niños mexicanos tienen más errores debido a que son los únicos que no escriben con letra cursiva, ya que ésta parece que ayuda a los otros niños a identificar mejor el principio y término de las palabras.

3. Con los niños que no dominan la segmentación, es notorio en sus textos que conforme más escriben, aparecen cada vez más estos dos fenómenos de segmentación.

4. Es frecuente que los niños peguen las palabras cuando una de ellas está compuesta por una o dos letras.
5. Las “mismas secuencias que producen la mayor parte de los problemas de hiposegmentación son también las que producen la mayor parte de los problemas de hipersegmentación” (Ferreiro: 1996, p. 70).
6. En los textos de los niños brasileños se registraron alrededor de 70 versiones diferentes del nombre Caperucita (*Chapeuzinho*). Algunas de estas versiones fueron consecuencia de que algunas letras comparten el mismo fonema, otras versiones derivaron de que los niños no modificaron la letra que iba antes del sufijo diminutivo, otras tantas fueron resultado de una hipersegmentación del sufijo, en otras versiones hubo “sustitución de ciertos grafemas por analogía gráfica con otros” (Ferreiro: 1996, p. 95), otras versiones surgieron debido a que los niños usaron acentos cuando realmente la palabra no los requiere, y algunas otras versiones fueron resultado de la mezcla de todo lo anterior. Algunos ejemplos de estas cinco explicaciones son:
    a. Fonética: Xapeusio
    b. Gramatical: Chapelzinho
    c. Semántica: Chapel zinho
    d. Analogía: Chapelzilho
    e. Gráfica: Chapéuzinho
7. En portugués el nombre de la protagonista está en masculino, tanto el sustantivo *Chapeuzinho* (Caperucito) como su adjetivo *Vermelho* (Rojo), pero al ser éste el nombre de una niña hubo algunos errores a nivel morfológico debido a que los niños no sabían cómo hacer la concordancia de género entre el nombre y la niña, por lo que se encontraron ejemplos como: Chapeuzinh**a** Vermelh**o** (Caperucita Rojo), Chapeuzinh**o** Vermelh**a** (Caperucito Roja), o Chapeuzinh**a** Vermelh**a** (Caperucita Roja).

En italiano el nombre de la protagonista también está en masculino (Cappuccetto Rosso) pero los niños italianos casi no tuvieron problemas en relacionar un nombre masculino con una niña, por lo que prácticamente no hay errores morfológicos ni ortográficos en relación con este nombre.

Con respecto a la puntuación, el análisis reveló los siguientes puntos:

8. En los textos italianos hay un mayor uso de los dos puntos y de las comillas que en los textos en español.
9. “los textos en español utilizan el doble de guiones de apertura/cierre que los textos en italiano”.
10. En los textos de los cuatro países “no hay casi diferencia en la frecuencia de uso de admiraciones e interrogaciones” (Ferreiro: 1996, p. 141).

2.4.2. Corpus Excale de escritura

El Instituto Nacional para la Evaluación de la Educación (INEE) es la organización mexicana encargada de ofrecer a las autoridades educativas y al sector privado herramientas idóneas para la evaluación de los sistemas educativos de nivel básico (preescolar, primaria y secundaria) y media superior. Con ello se pretende mantener un informe actualizado de la calidad educativa para trabajar en el mejoramiento de ésta, ya sea con el desarrollo de nuevos programas didácticos, el desarrollo de otras investigaciones orientadas al análisis lingüístico, dialectal o cultural, etc.

Como parte de este objetivo general el INEE desarrolló un corpus de textos escritos para evaluar: a) los diferentes tipos de discurso que utilizan niños y adolescente, b) la creatividad con la que se desenvuelven en sus textos y c) la calidad ortográfica.

Así, se elaboró el Corpus Excale de Escritura con los textos que se obtuvieron en la evaluación de los Exámenes de la Calidad y el Logro Educativo (Excale) de Español que aplica el INEE.

En la evaluación del Excale las preguntas varían en complejidad y temática dependiendo del grado escolar, pero en general es una prueba amigable que se preocupa porque los alumnos entiendan a la perfección las instrucciones, para así lograr que proporcionen respuestas descriptivas y argumentativas, con lo cual se evalúa el aprendizaje complejo y la creatividad de los estudiantes.

Es importante mencionar también que el Excale incluye para su evaluación un examen socioeconómico para así tener en cuenta el ambiente en el que se desenvuelve el niño, ya que como se ha mencionado anteriormente esto influye en gran medida en su aprendizaje y desarrollo.

Actualmente el Corpus Excale de Escritura está conformado por 14,314 textos de los cuales 3,768 pertenecen a niños de tercero de primaria, 4,848 a niños de sexto de primaria y 5,698 a jóvenes de tercero de secundaria.

Este corpus está a disposición de cualquier usuario en la página del INEE[16] y para su consulta la página ofrece una serie de herramientas que nos permite hacer tanto una búsqueda general como una muy detallada, como se podrá ver en la siguiente imagen.

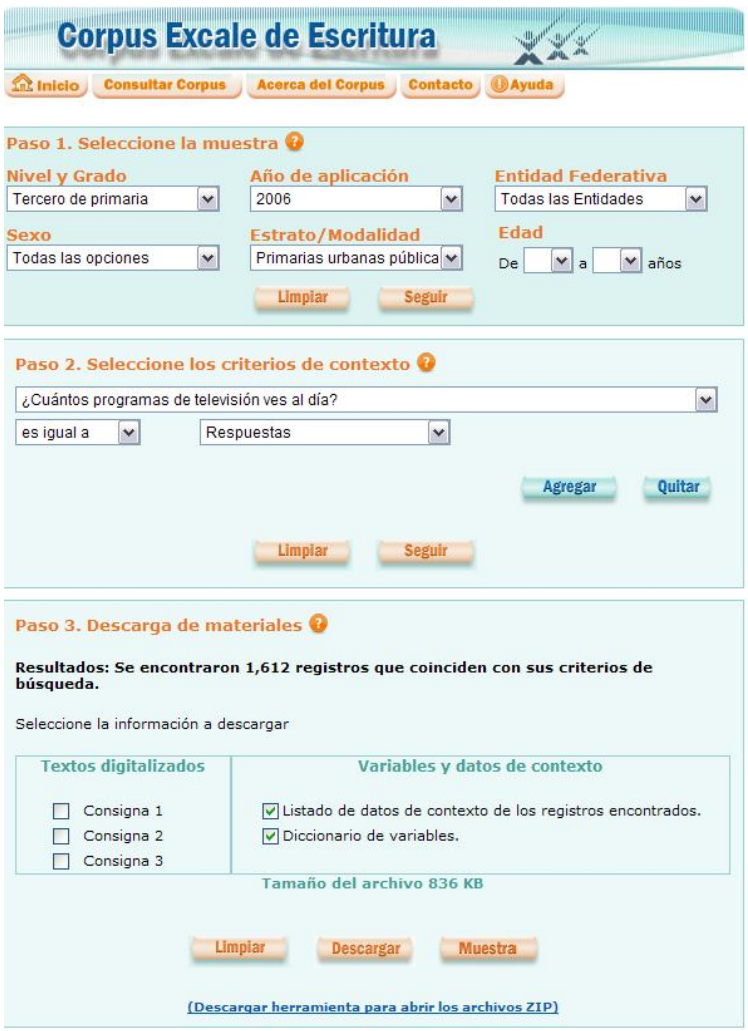


Ilustración A Corpus Excale de Escritura

En la imagen A podemos ver que para consultar los textos hay que seguir tres pasos: En el primero se solicita el nivel y grado escolar que se desea consultar, el año de la aplicación del examen, la entidad federativa, el género de los niños, el estrato social (urbano o rural), modalidad educativa (pública o privada) y la edad de los niños.

En el segundo paso se solicita los criterios de contexto, es decir, el tema sobre lo que tratan los textos; para ello el link despliega una amplia lista de temas donde se escoge uno solo. Algunos de estos temas son preguntas muy específicas que se

[16] http://www.inee.edu.mx/index.php?option=com_wrapper&view=wrapper&Itemid=1318

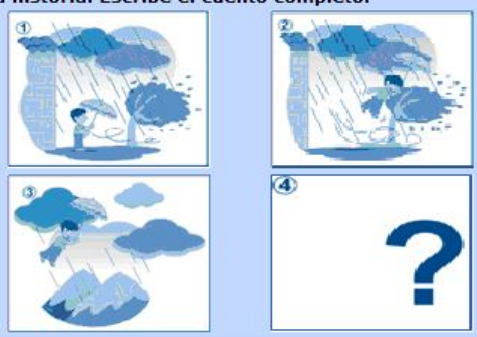
responden con un “sí” o un “no”, para estos casos los siguientes links son muy útiles, ya que se puede seleccionar si deseamos que las respuestas sean *iguales* o *diferentes* a un “sí” o un “no”.

En el tercer paso podemos ver el número de textos que cumple con los criterios que seleccionamos, y se nos dan las opciones para ver los textos o descargarlos.

Una vez que se abrieron los resultados podemos ver cada uno de los textos con su información correspondiente. Como se podrá ver en la imagen B esta información es muy clara, y aunque a primera vista no se entienda la columna de preguntas y respuestas, al acercar el “mause” a alguna pregunta se abrirá una ventana en donde se desglose la pregunta claramente. Algunas de estas preguntas son: ¿qué lengua aprendiste a hablar primero?, ¿sabe leer y escribir tu mamá?, ¿cuántas veces fuiste al cine en este año escolar?, etc.

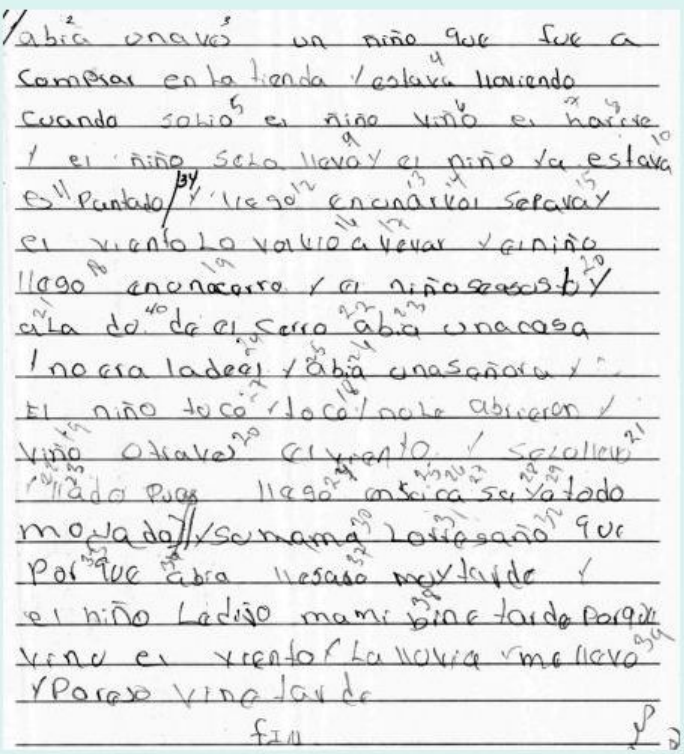
Consigna

Observa con atención las imágenes e imagina un final creativo para la historia. Escribe el cuento completo.



<< - < - > - >>

Imagen de Texto (22) Archivo de alta resolución



Había una vez un niño que fue a comprar en la tienda y estaba lloviendo cuando salió el niño vino el hombre y el niño se lleva y el niño ya estaba espantado y llego encontró se lavay el viento lo volvió a llevar y el niño llego encontró y el niño se asustó y ya la casa de el cerro había una casa y no era la de el y había una señora y el niño tocó y tocó y no le abrieron y vino otra vez el viento y se llevó llegado pues llego encontró se ya todo mojado y su mamá lo dijo señor que por que abra llegado muy tarde y el niño le dijo mami voy tarde porque vino el viento la lluvia y me llevo y pero vino tarde

FIN

Nota: las marcas que aparecen en los textos de los alumnos son el resultado del proceso de calificación de respuestas.Para mayor información sobre este sistema se sugiere refereirse a los siguientes textos: Referencia 1 , Referencia 2

<< - < - > - >>

Cerrar

METADATOS

Nivel y Grado:

Tercero de primaria

Edad: 10 **Año:** 2006

Entidad:

Chiapas

Sexo:

Hombre

Modalidad:

Primarias urbanas públi

Folio:

11130

Preguntas - Respuestas

ap001	-	1
ap002	-	1
ap008	-	2
ap012	-	3
ap023	-	0
ap024	-	4
ap052	-	1
ap053	-	1
ap054	-	0
ap055	-	5
ap056	-	4
ap057	-	2
ap058	-	0
ap069	-	3
ap070	-	0
ap071	-	1
ap072	-	0
ap073	-	1
ap074	-	2
ap076	-	0

Ilustración B Texto del Excale

Aunque este corpus puede servir como material de estudio a varios investigadores interesados en el discurso escrito de niños y adolescentes, el INEE debe realizar sus propios análisis para poder hacer una evaluación de la educación en México. Así, de la evaluación que se llevó a cabo en el 2008 se obtuvieron los siguientes resultados[17]:

- ✓ En promedio, por cada cien palabras, los alumnos de tercero de primaria cometen 31 errores ortográficos, los de sexto de primaria cometen poco más de 18 y los de tercero de secundaria cometen alrededor de 13 errores.
- ✓ Con respecto a la educación primaria, la modalidad con mayor índice de errores ortográficos es la rural pública y en relación con la educación secundaria es la telesecundaria.
- ✓ En general las escuelas privadas en todas las modalidades tienen el menor índice de error en comparación con las escuelas públicas.
- ✓ Los hombres presentan más errores que las mujeres, y los alumnos de extra-edad más que los de edad normativa[18].
- ✓ Tanto los grados de primaria como el de secundaría en todas las modalidades presentan mayor índice de error con la acentuación.
- ✓ Las palabras en las que más se equivocan todos los alumnos son las de uso más común, como por ejemplo las conjugaciones de los verbos ser, estar e ir.

Con todo esto se concluyó que la frecuencia de error en los alumnos es muy alto y se propone una técnica didáctica que consiste en que el maestro muestre al alumno el cambio de significado que se produce con un error ortográfico, por ejemplo entre *el transito* y *él transitó.*

También se sugiere a) que los maestros no dejen de lado el análisis de la ortografía y la revisión de los escritos de los niños para corregirlos y b) que los maestros deben enseñar a los alumnos a planear el texto, después producirlo y finalmente revisarlo.

---

17 http://www.inee.edu.mx/images/stories/Publicaciones/Reportes_investigacion/Ortografia/Partes/erroresortograficos06.pdf

18 Los alumnos de "extra-edad" son aquellos que son grandes para el grado en el que están, por ejemplo un niño de 8 años en un grupo de 1º año donde la edad "normativa" es de 6 años.

2.4.3. CHILDES

CHILDES debe sus siglas a su nombre en inglés Child Language Date Exchange System y como su nombre lo indica, CHILDES es un sistema de cómputo de intercambio de datos.

Este sistema fue creado por Brian MacWhinney y Catherine Snow con la finalidad de recopilar una amplia base de datos sobre lenguaje infantil espontáneo. De esta forma CHILDES se ha estado desarrollando desde 1984 logrando obtener hasta la fecha 130 corpuora diferentes entre los cuales podemos encontrar tanto material transcrito, como material en audio y en video, en 20 idiomas distintos.

Estos 130 corpora fueron divididos en 6 categorías:

- Inglés
- No- Inglés
- Narrativos
- Libros
- Problemas de lenguaje
- Adquisición bilingüe

CHILDES está compuesto por tres programas que en un principio fueron diseñados por distintas personas para cumplir diferentes propósitos, pero con el paso del tiempo estos tres programas se complementaron para formar uno solo: CHILDES, el cual ya forma parte del TalkBank[19]. Por tanto, ahora no sólo contamos con el material que nos proporciona CHILDES, sino que también podemos contar con el material de otros seis sistemas[20] que también forman parte del TalkBank.

De esta forma CHILDES está compuesto por:

[19] El TalkBank es un proyecto de investigación interdisciplinaria desarrollado entre los años de 1994 y 2004 por la Universidad Carnegie Mellon y la Universidad de Pensilvania bajo el financiamiento de la Fundación Nacional de Ciencia. El TalkBank tiene como propósito recolectar investigaciones que se hayan servido de herramientas computacionales para estudiar la comunicación entre humanos y entre animales. Para más información se puede consultar la página http://talkbank.org/.

[20] AphasiaBank, BilingBank, CABank, DementiaBank, PhonBank y TBIBank.

- CHILDES: Es el sistema que contiene la recopilación de los corpora.
- CHAT: El Codes for the Human Analysis of Transcripts (Códigos para el Análisis y Transcripciones Manuales) es un sistema formulado por varios investigadores entre los años de 1984 y 1988. La finalidad del CHAT fue integrar códigos de transcripción y codificación que se pudieran aplicar a cualquier tipo de corpus sin importar el material o el idioma en que esté registrado.
- CLAN: El Computerized Language Analysis (Análisis de Lenguaje computarizado) es un programa creado por Leonid Spektor y se encarga de analizar datos en los siguientes niveles:

    - Léxico
    - Morfológico
    - Sintáctico
    - Discursivo
    - Fonológico

Pero para llevar a cabo estos análisis el CLAN utiliza los siguientes comandos:

- FREQ: Frecuencia
- KWAL: Analizador de ítems específicos (nombre, pronombre, adjetivos, verbos, etc.)
- LEX: Analizador léxico
- MOR: Analizador morfológico
- COMBO y COOCCUR: Analizadores sintácticos
- CED: Código de edición, se emplea para realizar análisis a nivel discursivo
- PHO: Analizador fonológico

Aunque CHILDES no es un proyecto de investigación sobre la adquisición de la lengua escrita como los proyectos que mencionamos en los puntos anteriores y como del que se deriva esta tesis; la recopilación que ha llevado a cabo CHILDES y las herramientas que

ofrece de forma gratuita para el análisis de ese material han propiciado hasta la fecha la publicación de más de 3000 estudios sobre el lenguaje infantil[21].

Por tanto, más que exponer en este apartado el desarrollo y resultado de un proyecto, mostramos una fuente muy importante de materiales para que se desarrollen muchos más proyectos.

[21] La bibliografía de algunas de las publicaciones se pueden consultar en http://talkbank.org/usage/childesbib.pdf

3. BASES METODOLÓGICAS DE CORPUS LINGÜÍSTICOS

Como vimos en el capítulo anterior, las dos investigaciones que expusimos requirieron de la recopilación de textos para llevar a cabo sus análisis. Asimismo, CHILDES al lograr recopilar una enorme cantidad de datos se convirtió en una gran fuente de información para muchos investigadores.

Es así como esa recopilación de textos, llamada corpus, es una herramienta muy importante, incluso en ocasiones es obligatoria para llevar a cabo muchas investigaciones relacionadas con el lenguaje. De esta forma en este capítulo definiremos en detalle lo que es un corpus, las consideraciones que hay que tener en cuenta para su recopilación, y su relación con la tecnología computacional.

3.1. Corpus lingüístico

3.1.1. Descripción de corpus lingüístico

Desde un punto de vista etimológico podemos definir al **corpus** como un conjunto de textos, por otro lado la Real Academia Española lo define como un "conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación".

Ambas definiciones son muy limitadas, por lo que una definición más detallada sería que un **corpus lingüístico** "consiste en la recopilación de un conjunto de textos de materiales escritos y/o hablados, agrupados bajo un conjunto de criterios mínimos, para realizar ciertos análisis lingüísticos" (Sierra: 2008, pág. 445). ¿Qué significa esto? Los corpus pueden constituirse relativamente por cualquier texto escrito o hablado, eso quiere decir que podemos encontrar corpus compuestos básicamente de cualquier material, ya sean cartas, periódicos, revistas, libros, grabaciones de audio, videos, etc. Pero no por ello significa que la recopilación de los corpus es desmedida y arbitraria, al contrario, otra de las características del corpus es que éste sea planeado y construido con la cantidad de textos necesarios para llevar a cabo el análisis del aspecto lingüístico que se quiera estudiar. Y aquí tocamos otro punto muy importante, con ayuda de los corpus se pretende analizar sólo aspectos lingüísticos de la lengua, esto significa que los

corpus, más que registros de textos, son representaciones de la lengua, pero no de la totalidad de ella, sino sólo de una parte.

Como complemento a todo esto, en la actualidad los corpus están muy relacionados con algunas herramientas computacionales, por lo que se puede considerar una sub-categoría de éstos con el nombre de **corpus informatizado**, el cual, además de tener las características ya mencionadas, se define como un conjunto "de textos elegidos y anotados con ciertas normas y criterios para el análisis lingüístico, de forma que se sirve de la tecnología y de las herramientas computacionales para generar resultados más exactos" (Sierra: 2008, pág. 455).

Estos corpus informatizados revolucionaron los métodos de investigación ya que no sólo aportan resultados más exactos, sino que además contribuyen a que el análisis se lleve a cabo de forma más fácil y rápida. Por tanto, los corpus dejaron de ser exclusivos de la lingüística y se fueron relacionando con muchas otras ciencias, lo cual se manifiesta con los diccionarios, los traductores, las bases de datos, la inteligencia artificial, etc.

Derivado de este crecimiento y éxito de los corpus informatizados se puede considerar una disciplina más en la rama de la lingüística: la **lingüística de corpus**, la cual consiste en analizar grandes y pequeñas cantidades de información con ayuda de las herramientas computacionales.

3.1.2. Compilación de corpus

Como acabamos de mencionar, a pesar de que los corpus pueden estar constituidos por una gran variedad de textos, no significa que se pueda recolectar cualquier cosa que tenga letras y/o cualquier cosa que reproduzca palabras. Los textos que conforman un corpus tienen que cumplir con ciertos criterios para que con base en el análisis de éstos los resultados que se arrojen sean confiables.

Por supuesto algunos criterios para la compilación de textos pueden variar en cada investigación, pero existen algunas características generales que se deben tomar en cuenta para la elaboración de cualquier corpus sin importan la disciplina en la que se desarrolle (Sierra: 2008).

Lo primero que se debe de tomar en cuenta es que los textos tienen que ser **representativos** de la lengua o las lenguas que se quieren estudiar, para ello hay que

ser muy selectivos respecto a la zona geográfica, la cultura, la etnia, la época etc., ya que de esta forma se podrá cumplir con el objetivo que se plantee.

Pero tener en cuenta eso no es suficiente, hay que estar conscientes también de algunos otros detalles para estar satisfechos con la recolección que se haga:

- Se debe de tener claro que tan general o que tan especifico será el corpus, ya que con base en eso se podrán recolectar textos representativos de todo un país, de un estado, de un municipio, de una región, de un grupo social, etc.
- Es importante tener cuidado de que los textos provengan de personas que realmente representen el aspecto lingüístico que se quiere estudiar, por ejemplo si se hiciera un análisis de las diferencias fonológicas del portugués de Brasil y de Portugal es recomendable que los textos deriven de personas nativas de estos países cuya lengua materna fue por supuesto el portugués, y se descartarían extranjeros (aunque tengan años viviendo en cualquiera de esos países), así como hablantes nativos que han pasado demasiados años en otro país.
- Sin importar la especificidad del corpus se tienen que registrar todos los detalles sobre la zona geográfica de donde se extrajo. De la misma forma es necesario tener registro de algunos datos de las personas de quienes se obtuvieron los textos, como edad, género, extracto social, nivel de estudios, etc. Y también hay que registrar el año o la época representada por el corpus.

Esto hay que hacerlo debido a que muchas veces esta información es relevante y decisiva tanto para la propia investigación como para otros estudios que quieran consultar el material. Por tanto, es muy importante que esta información se exhiba junto con el corpus.

Otra característica importante para que el corpus sea representativo es que necesita tener *variedad y equilibrio* de textos, esto significa que los textos que conforman el corpus tienen que representar diferentes rubros de forma equilibrada, es decir, retomando nuestro ejemplo anterior sobre Brasil y Portugal, para que el corpus sea variado los textos deben provenir de diferentes regiones de cada uno de estos países, y por cada una de estas regiones se debe recolectar la misma cantidad de textos para que el corpus sea equilibrado. En caso de que el corpus sea más específico se tienen que considerar rubros diferentes, por ejemplo, si se pretende hacer el estudio con material oral que se obtendrá de una sola persona en una semana, los rubros pueden ser

temporales, así se dividiría día uno, día dos, etc. y se obtendría la misma cantidad de textos en cada uno de estos días.

Una vez que se ha reflexionado sobre la representatividad del corpus hay que considerar el **tamaño** de éste. Tener esto en cuenta es importante debido a que muchas veces se cae en el error de que entre más textos el corpus dará mayores resultados, pero lo que en realidad acontece es que el análisis se vuelve extenuante y más que una gran cantidad de resultados, lo que arroja el corpus es una gran cantidad de datos vacios que no aportan nada de información para la investigación. Así, aunque se tenga en mente hacer un corpus que abarque muchos rubros hay que hacer una recolección proporcional de textos en cada uno de ellos, recordemos que un corpus es la representación de una parte de la lengua y no de la totalidad de ella, por lo que hay que cuidar en recolectar los textos donde se muestre ese fenómeno lingüístico que se quiere estudiar (Sierra: 2008).

Considerando las ventajas que las herramientas computacionales ofrecen, en la actualidad se trabaja básicamente con los corpus informatizados, a tal grado que de muchos de los corpus más importantes, que en su tiempo se hicieron a mano, ya se han guardado versiones en formatos electrónicos. Por tanto, debido a que hoy en día resulta absurdo hacer un corpus para analizarlo manualmente, es importante guardar la recolección que se haya hecho en una computadora.

Las ventajas de hacer esto son varias:

- ✓ Los textos pueden ser ordenados y clasificados manteniendo una limpieza en el corpus.
- ✓ En el internet se pueden encontrar suficientes textos para constituir un corpus o se pueden encontrar varios corpus ya constituidos que otros investigadores comparten. Mucho de este material se puede encontrar de manera gratuita, por lo que además de ahorrarse dinero, se ahorra tiempo de recolección.
- ✓ Se pueden hacer conteos exactos de forma automática.
- ✓ Si se desea analizar algunos elementos en especial, la computadora puede extraer todos los que encuentre en el corpus de forma rápida y precisa, de esta forma el investigador sólo se ocupa en analizarlos y se ahorra el tiempo de estarlos buscando con relecturas de todo el corpus.

- ✓ El corpus puede guardarse en dispositivos de almacenamiento masivo para ser transportado a cualquier lado y poder ser consultado en cualquier computadora compatible con el programa en que fue guardado[22].
- ✓ Las fichas se pueden actualizar al mismo tiempo sin necesidad de que sean revisadas una por una.

A pesar de estas ventajas, constituir un corpus con soporte electrónico y hacer su análisis con herramientas computacionales no es tarea sencilla, ya que hay que llevar a cabo otras actividades y tener en cuenta algunas consideraciones (Sierra: 2008):

- ✖ Guardar un corpus escrito en formato electrónico se puede hacer de dos maneras: transcribiendo o digitalizando los textos. El primero consiste en copiar manualmente los textos impresos a algún formato electrónico de escritura[23]; y el segundo consiste en obtener imágenes electrónicas de los textos. Este último método, a diferencia del primero, no sólo requiere del uso de una computadora, sino también necesita de otro aparato llamado escáner.

Con respecto a los corpus orales, si la recolección de textos se llevó a cabo con un aparato tradicional como grabadora o videograbadora de casete, o con algún otro aparato que no puede ser conectado a una computadora, pasar esos textos a formato electrónico requerirá de otro u otros aparatos que podrían ser costosos dependiendo de la calidad en como se quiera guardar los textos.

Sea cual sea el método que se seleccione, dependiendo del tamaño del corpus y de la cantidad de personas que trabajen con él, el proceso para obtener un corpus con formato electrónico puede requerir varias horas de trabajo, o incluso hasta días, semanas, meses o años.

- ✖ Varios formatos electrónicos cuentan con algunas herramientas fáciles de utilizar para analizar rápidamente algunos elementos, pero los análisis que se puedan hacer con estas herramientas son muy básicos. Si se quiere herramientas computacionales para llevar a cabo análisis más específicos, como por ejemplo

[22] Actualmente la mayoría de los programas básicos que existen son compatibles en todas las computadoras.
[23] TXT, Word, Excel, Access, etc.

de todas las frases nominales o de un verbo en particular con todas sus conjugaciones, etc., se necesita un software determinado.

Existen software ya diseñados que se pueden encontrar en la red, algunos son gratuitos y otros pueden tener varios precios, pero para el uso de cualquiera de éstos es necesario que el corpus cumpla con ciertas características, las cuales varían dependiendo del software que se use.

En caso de que el corpus no cumpla con las características o simplemente no se haya encontrado el software adecuado existe la posibilidad de crear uno que se adapte a las características del corpus y a las necesidades de la investigación.

Cualquiera que sea la decisión es necesario tener o adquirir un conocimiento de ese software o contar con la ayuda de un especialista en computación.

- Para investigaciones pequeñas muchas veces es suficiente contar simplemente con una computadora, pero si la investigación es más compleja se necesita contar con varios equipos de cómputo, lo cual puede llegar a ser muy costoso.
- Conforme la tecnología vaya avanzando se tienen que llevar a cabo las actualizaciones pertinentes, las cuales pueden consistir desde la sustitución del equipo de cómputo hasta la modificación del software y todos los elementos que lo conforman (formatos, etiquetas, códigos, etc.). Por supuesto esto es necesario sólo si se quiere que el corpus siga siendo vigente y adaptable.
- A pesar de que la tecnología cada día es más sofisticada aún puede llegar a haber ciertas fallas técnicas, por lo que es importante tener un respaldo que proteja el trabajo que se ha hecho, ya que sería muy frustrante perder en unos segundos todo lo que se logró obtener en meses.

Finalmente, algo que también es muy importante tener en cuenta para la recolección son los permisos correspondientes para poder hacer uso de los textos (Sierra: 2008).

En la actualidad existen leyes internacionales que protegen los derechos de propiedad intelectual de los autores. Por tanto, si el corpus está constituido por material que otra u otras personas produjeron (ya sea de manera oral o escrita), por ley es necesario realizar algunos procedimientos y exponer cierta información junto con el corpus.

En caso de que el material se haya obtenido directamente con el hablante por medio de testimonios, charlas, entrevistas, chat, etc., es preciso solicitarle permiso para publicar el material como parte de la investigación. Asimismo, si el material se obtuvo de una o varias publicaciones ya hechas, en algunos casos también es necesario solicitar permiso de los titulares de la obra.

Posiblemente para algunos estudios pequeños es suficiente que los permisos se otorguen de manera oral, pero si el corpus pertenecerá a un proyecto más formal, por ley se requiere que los permisos de los titulares o entrevistados se hagan por escrito antes de llevar a cabo el análisis sobre los textos.

Un caso donde no es necesario obtener estos permisos es cuando el estudio se está llevando a cabo con fines no lucrosos y que sólo exhibe fragmentos del material en caso de ser publicado.

Pero independientemente de los permisos es indispensable dar la bibliografía o referencia al origen de los textos junto con el corpus.

De la misma manera, una vez terminada la investigación que se hizo con el corpus, ésta también está protegida como propiedad intelectual, por lo que, en caso de que haya habido más personas involucradas, es importante darles crédito por su trabajo, así como agradecer la ayuda a los patrocinadores.

3.2. Codificación de corpus

Como hemos descrito en el punto anterior, a pesar de las dificultades para trabajar corpus informatizados aún se pueden considerar más ventajosos que los corpus manuales, y no sólo por todo lo que ya se mencionó, sino porque además conforme avanza la tecnología hay aparatos, técnicas, herramientas, etc., que se quedan en el olvido y de la misma forma tarde o temprano los corpus manuales caducarán, de hecho, actualmente se trabaja ya tanto con corpus informatizados, que al hablar de corpus básicamente ya queda por entendido que se habla de los informatizados como si no existiera ningún otro.

Por tanto, para que un corpus perdure por muchos años es prácticamente obligatorio que se haga o se traspase a formatos con soporte electrónico para ser trabajados por algún software.

Dependiendo del tipo de discurso que se quiera analizar, las herramientas computacionales varían considerablemente, pero debido a que el corpus recolectado por

el proyecto *Aprender mientras se enseña: una experiencia de acompañamiento en la enseñanza de la Lengua escrita* es escrito y en consecuencia el objeto de estudio de esta tesis también lo es, en este apartado nos centraremos en el etiquetado y lenguajes de marcado diseñados específicamente para trabajar con corpus escritos.

Asimismo, explicaremos algunos lenguajes de marcado[24], algunos programas que hacen análisis en diferentes niveles lingüísticos y finalmente describiremos a detalle lo que es el XML, debido a que éste fue el programa utilizado en el proyecto *Aprender mientras se enseña: una experiencia de acompañamiento en la enseñanza de la Lengua escrita*.

3.2.1. Etiquetado

Una vez que se ha logrado obtener el corpus con soporte electrónico, una característica muy importante que debe de tener es la limpieza, es decir, no debe haber sido tratado aún por ningún lenguaje de marcado.

Cuando el corpus está limpio se puede etiquetar de distintas maneras, es decir que se le pueden colocar marcas que nos ayuden a identificar y describir los elementos de interés para el análisis. Para ello se puede diseñar un repertorio de etiquetas o se puede utilizar algún repertorio de los varios que ya existen. Sea cual sea la decisión que se tome para etiquetar el corpus existen siete normas que según Leech (Pérez: 2002, p. 51) deben aplicarse en la codificación de éstos:

1) Debe ser posible eliminar las etiquetas añadidas a un texto anotado y recuperar el texto original sin que éste sufra modificación alguna.
2) Debería ser posible también extraer las anotaciones de los textos y almacenarlas de forma independiente, por ejemplo en una base de datos relacional o en líneas paralelas al texto original.
3) El sistema de anotación usado debe estar basado en unas directrices, documentadas y accesibles al usuario final del corpus, de modo que pueda tener acceso tanto a un listado completo de las etiquetas usadas como a las decisiones tomadas en el proceso de etiquetación.
4) Debe ser posible incluir información sobre la autoría de la codificación del texto, de forma que sea posible saber si se ha realizado manualmente (y por quién), o si se ha realizado de forma automática con o sin revisión posterior por un lingüista.
5) Se debe hacer al usuario final consciente de que las anotaciones añadidas al corpus no son infalibles, sino que simplemente constituyen

[24] O también llamados metalenguajes. Éstos son lenguajes que tratan o describen a otro lenguaje, en este caso los lenguajes de marcado se sirven de las etiquetas para hablar o describir el lenguaje del corpus.

una herramienta de ayuda para el análisis. Cualquier anotación que se añada al corpus será, por definición, un acto de interpretación y de análisis del texto, por lo que es susceptible de incorrecciones e inexactitudes.

6) Los sistemas de anotación han de estar basados en la medida de lo posible en principios teóricamente neutrales y sobre los que exista un acuerdo amplio en el seno de la comunidad científica.
7) Ningún sistema de anotación posee, a priori, el derecho de ser considerado estándar. Los estándares, cuando existen, se desarrollan por el consenso de los usuarios, como fue el caso del sistema de referencia COCOA, muy usado hace unos años o de los estándares propuestos por TEI, usados actualmente en la mayoría de los proyectos.

Para diseñar un repertorio de etiquetas hay que servirse de un lenguaje de marcado y dentro de todos los que existen, el más reconocido es el Standard Generalized Markup Language (SGML). Éste es un lenguaje de marcas textuales que surgió en 1986 con el nombre de GML y fue aprobado como estándar de codificación en el mismo año por la International Organization for Standardization (ISO).

Con el SGML se pueden diseñar etiquetas para marcar y describir cualquier elemento de un texto. De este lenguaje se desprenden otros dos que al ser más específicos son menos complejos y más fáciles de utilizar por personas que no son especialistas en computación. Estos otros lenguajes son: HTML y XML, derivándose de este último los lenguajes: XHTML, Wireless ML, GladeXML, MXML, XAML, XForms, XUL y XBL.

Estos son algunos de los lenguajes más populares y sencillos de utilizar, pero por supuesto se pueden encontrar muchos otros que cumplan con las necesidades y expectativas de diversas investigaciones.

3.2.2. Tipos de anotación

Como acabamos de ver hay varios y diferentes lenguajes que pueden ayudar a diseñar un repertorio de etiquetas para codificar en un corpus elementos determinados, pero como también hemos mencionado, existen varios proyectos que han desarrollado estándares de anotación y ofrecen repertorios de etiquetas que pueden ayudar a llevar a cabo análisis lingüísticos en distintos niveles.

Al codificar los corpus con alguno de estos proyectos los investigadores se ahorran tiempo en definir las etiquetas y diseñar el programa que los ayudará a etiquetar, ya que además del repertorio de etiquetas muchos de estos proyectos ofrecen las herramientas computacionales que ayudan a codificar el corpus (Sierra: 2008). Por tanto, los investigadores que se sirven de estos proyectos dedican poco tiempo para codificar y más tiempo para analizar.

A continuación describiremos los tipos de anotación y algunos proyectos cuyas etiquetas se centran en algún nivel lingüístico determinado.

3.2.2.1. Anotación morfosintáctica

La anotación morfosintáctica es considerada como POST, que en inglés significa Part Of Speech Tagging y consiste en describir la naturaleza y función de una palabra en una cadena nominativa (Sierra y Rosas: en prensa).

Dentro de todos los proyectos que trabajan con POST, el más conocido es el EAGLES.

Debido a la cantidad de etiquetas que fueron surgiendo de una gran diversidad de corpus, en 1993 la Dirección General XIII de la Comisión Europea puso en marcha el proyecto Expert Advisory Group of Language Engineering Standards (EAGLES) con el propósito de "elaborar, mediante un amplio consenso, recomendaciones y especificaciones para áreas concretas de la tecnología lingüística a partir de los resultados de trabajos en curso en diversas organizaciones del ámbito comunitario y promover su adopción en futuros proyectos" (Arrearte: 1999, p. 41).

Gracias a la gran variedad de recomendaciones y especificaciones que diseñó el EAGLES ha habido ya varios proyectos de investigación que se basaron en él, algunos para llevar a cabo sus análisis con las etiquetas que ofrece, y otros que con base en éstas desarrollaron adaptaciones en varios idiomas.

Para el castellano tenemos las adpataciones de las etiquetas EAGLES en un analizador morfológico que fue desarrollado por el Departamento de Lenguajes y Sistemas Informáticos de la Universidad Politécnica de Cataluña. Un ejemplo de estas adaptaciones son:

Forma	Etiqueta	Significado
bonitas	AQ0FP00	A(adjetivo) Q(Calificativo) 0(no grado) F(femenino) P(plural) 0(no caso) 0(no participio)
el	TDMS0	T(artículo) D(definido) M(masculino) S(singular) 0 (no caso)
oyente	NCCS000	N(nombre) C (común) C(común(neutro)) S(singular) 0(no caso) 0(no género semántico) 0(no apreciativo)
cantamos	VMIP1P0	V (verbo) M(principal) I(indicativo) P(presente) 1(primera persona) P(plural) 0 (no género)

3.2.2.2. Anotación sintáctica

La anotación sintáctica consiste en describir la relación funcional que se produce entre las palabras de una oración, es decir, se hace un análisis de la oración, la cual es conocida como *parsing* (Sierra y Rosas: en prensa).

El parsing se puede dividir en dos niveles: parcial y total. El primero es un análisis de los constituyentes únicamente, por ejemplo:

O: (Sujeto= [La (art) protagonista (sus) de (prep) la (art) película (sus)] Complemento= [estaba (vrb) casada (part) con (prep) el (art) malo (adj)])

En cambio el parsing total es un análisis de los constituyentes y su relación sintáctica, por ejemplo:

O: (Sujeto= [FN= La (art.) protagonista (sus) CAdn= [de (prep) FN= [la (art) película (sus)]]] Complemento= [FV= [estaba (vrb) casada (part)] CC= [con (prep) FN= [el (art) malo (adj)]]])

Uno de los proyectos que hace parsing total es el Penn Treebank[25], el cual se llevó a cabo en el Departamento de Computación y Ciencia de la Información de la Universidad de Pensilvania. Este proyecto diseñó un programa que al recibir cualquier oración por parte del usuario elabora un árbol sintáctico con base en el análisis

[25] http://www.cis.upenn.edu/~treebank/

morfológico de la oración. Este programa se diseñó con el propósito de contar con una herramienta que ayude a la desambiguación de las oraciones. Por ejemplo, si tenemos una oración como *El hombre miró a la niña con el telescopio en el parque*, el Treebank elabora varios árboles sintácticos dependiendo de la cantidad de versiones de significado que tenga la oración[26]:

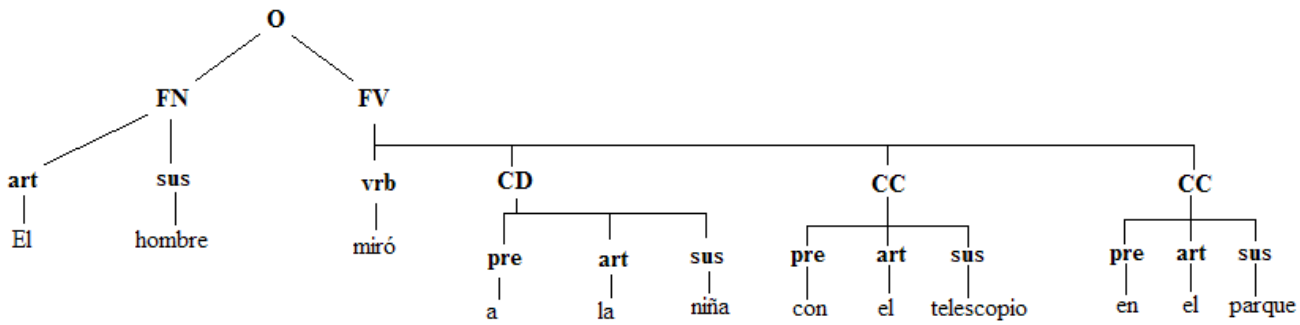


Ilustración C El hombre miró a la niña con el telescopio en el parque I

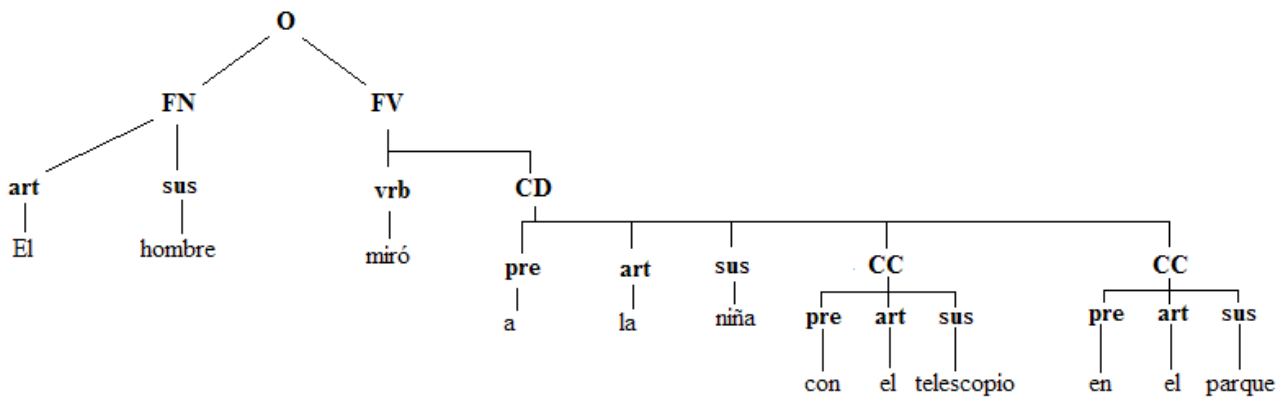


Ilustración D El hombre miró a la niña con el telescopio en el parque II

En el primer árbol se entiende que un hombre estaba en el parque y utilizó un telescopio para ver a una niña, y en el segundo árbol se entiende que un hombre vio una niña que tenía un telescopio y estaba en el parque. Por supuesto la oración que mostramos como ejemplo puede tener más de dos versiones de significado, pero estos árboles son suficientes para mostrar lo que hace el Penn Treebank.

[26] El Penn Treebank está en inglés, por tanto, los árboles que ponemos como ejemplo no se obtuvieron de él, pero muestran exactamente como éste desglosa las oraciones.

3.2.2.3. Anotación semántica

La anotación semántica consiste en describir el significado de una palabra de acuerdo con su propia naturaleza, con su posición dentro de una red semántica o con las relaciones léxicas (sinonimia, hiponimia, polisemia, etc.) que tenga con otras palabras (Sierra y Rosas: en prensa).

En la Universidad de Princeton se desarrolló una base léxica de datos en inglés llamado WordNet con el objetivo de desarrollar una herramienta que ayudara a la desambiguación, pero a diferencia del Penn Treebank, el WordNet no pretende desambiguar el significado de las oraciones, sino que con ayuda de un análisis semántico pretende desambiguar el significado y la función léxica de las palabras.

Para la elaboración del WordNet sólo se recopilaron sustantivos, verbos, adjetivos y adverbios, y aunque en un principio puede parecer que esta base de datos es un diccionario, en realidad el WordNet es una base de datos que no sólo nos ofrece todos los significados que pueda tener una palabra, sino que además nos ofrece también todas las relaciones léxicas que tiene. Por ejemplo veamos una consulta que se hizo en esta base de datos con la palabra *man* (hombre):

Noun[27]

- S: (n) **man**, adult male (an adult person who is male (as opposed to a woman)) *"there were two women and six men on the bus"*
- S: (n) homo, **man**, human being, human (any living or extinct member of the family Hominidae characterized by superior intelligence, articulate speech, and erect carriage)
    - *direct hyponym* / *full hyponym*
    - *member holonym*
        - S: (n) genus Homo (type genus of the family Hominidae)
            - *member holonym*
            - ***member meronym***
                - S: (n) homo, **man**, human being, human (any living or extinct member of the family Hominidae characterized by superior intelligence, articulate speech, and erect carriage)
                - S: (n) Homo sapiens sapiens, modern man (subspecies of Homo sapiens; includes all modern races)
            - *direct hypernym* / *inherited hypernym* / *sister term*
    - *domain term category*

[27] http://wordnetweb.princeton.edu/perl/webwn

- S: (n) **man** (a male subordinate) *"the chief stationed two men outside the building"; "he awaited word from his man in Havana"*

Verb

- S: (v) **man** (take charge of a certain job; occupy a certain work place) *"Mr. Smith manned the reception desk in the morning"*
- S: (v) **man** (provide with workers) *"We cannot man all the desks"; "Students were manning the booths"*

Para no ocupar mucho espacio se eliminaron varias definiciones que ofreció la base de datos, pero aún así podemos observar la variedad de definiciones y funciones léxicas que el WordNet nos da de una palabra. Asimismo, al seleccionar alguna de las definiciones la base de datos nos ofrece las relaciones léxicas que tiene con otras palabras, y al seleccionar alguna de estas otras palabras la base nos ofrecerá otras más y así sucesivamente, como se puede ver en el ejemplo a partir de la segunda definición.

3.3. Lenguaje de marcado XML

El XML, E**x**tensible **M**arkup **L**anguage (Lenguaje de Marcado Extensible), es un metalenguaje, es decir, es un lenguaje que ayuda a analizar otro lenguaje, en este caso el discurso infantil escrito.

Este lenguaje fue creado por la *World Wide Web Consortium* (W3C), y al igual que el HTML, el XML se basa en la norma internacional sobre información estructurada, en el SGML. Pero a pesar de sus semejanzas, estos dos metalenguajes tienen funciones diferentes: el HTML se encarga del aspecto de los datos y el XML del significado de éstos (Goldfarb y Prescod: 1999).

El objetivo del XML es ayudarnos a tener representaciones digitales de cualquier tipo de documento, con la finalidad de que la computadora pueda obtener de ellos lo que necesitamos, por supuesto de forma fácil, rápida y precisa.

Al igual que el SGML, el XML no es una entidad física como un documento con páginas, sino que es una entidad lógica compuesta por *elementos* (Arrearte: 1999) los cuales son anotados como etiquetas. Por ejemplo:

```
<musica>
        <barroca>
                <instrum>
                        <i tipo="cuerda"/>violín
```

```
                <i tipo="teclado"/>organo
                <i tipo="viento"/>trompeta
            </instrum>
            <compo>
                <c inst="organo"/>Johann Sebastian Bach
                <c inst="trompeta"/>Gottfried Reiche
                <c inst="violín"/>Arcangelo Corelli
            </compo>
        </barroca>
</musica>
```

En este ejemplo podemos ver que tenemos un documento que contiene el elemento "musica", el cual contiene el elemento "barroca" y éste a su vez contiene los elementos "instrum" y "compo", y así sucesivamente hasta llegar a donde está la información.

Es importante notar que cada uno de estos elementos está anotado en etiquetas, las cuales, en cualquier etiquetado XML, se pueden colocar de dos maneras:

- Una etiqueta de *apertura* por ejemplo "<musica>" siempre debe tener su etiqueta de *cierre* "</musica>" aunque haya más pares de etiquetas entre ellas.
- En lugar de usar pares de etiquetas también se pueden usar etiquetas solas, pero la diagonal que usan va del lado derecho, a diferencia de la etiqueta de *cierre*; por ejemplo, <i/> y <c/>.

Además de los elementos, las etiquetas pueden contener otros datos que llamamos *atributos* (Arrearte: 1999), los cuales dan información más detallada de lo que se está etiquetando, en nuestro ejemplo anterior tenemos los atributos *tipo=""* e *inst=""*; el primero especifica qué tipo de instrumento es el que está etiquetado y el segundo atributo especifica qué instrumento toca el compositor señalado.

Asimismo, como vemos en el ejemplo, el etiquetado puede organizarse con líneas separadas y sangrías, pero esta estructura sólo se hace para que sea fácil de leer por el ojo humano, es decir, el XML no necesita de esa estructura, él puede leer los datos sin las sangrías o incluso si toda la información estuviera encadenada en una sola línea.

A grandes rasgo hemos descrito cómo se etiqueta con el XML pero antes de llevar a cabo cualquier etiquetado se debe diseñar un *esquema* (schema/DTD), el cual es el formato donde el usuario tiene que declarar elementos y atributos, es decir, decide qué etiquetas necesita y bajo qué criterios, pero tiene que estar muy atento al hacer este

esquema y después al etiquetar, porque de acuerdo con las reglas que él mismo imponga en el esquema, se tiene que llevar a cabo el etiquetado del corpus.

Debido a que gran parte del etiquetado se hace a mano, para asegurarse que éste cumpla con reglas básicas, el XML cuenta con dos herramientas llamadas *Bien formado* y *Validación del archivo XML*: La primera revisa que las etiquetas estén bien y la segunda revisa que las reglas del esquema se estén llevando a cabo de acuerdo con las estructura que se le dio (Goldfarb y Prescod: 1999).

Finalmente, otra herramienta del XML, es el **XSL**, E**x**tensible **S**tyle **L**anguage (Lenguaje de Estilo Extensible), y su objetivo es presentar en un formato más comprensible el trabajo hecho en XML como se puede ver en las imágenes de abajo.

Ilustración E Formato XML

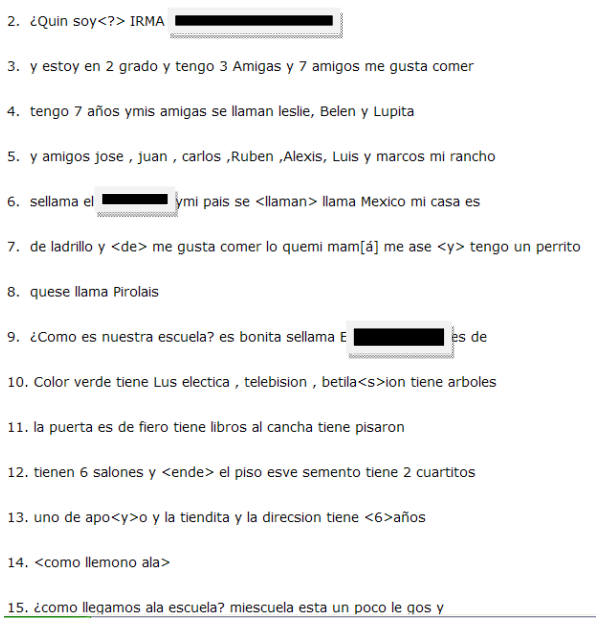


Ilustración F Formato XSL

En la imagen E podemos ver el corpus en formato XML, y en la imagen F se ve el mismo fragmento de la imagen anterior, sólo que en formato XSL[28].

Es importante mencionar, que al igual que el XML, el XSL también necesita que el usuario diseñe un esquema donde señale qué es lo que se quiere (tablas, líneas, dibujos, ventanas, enlaces, etc.) y cómo lo quiere (colores, número de filas, número de columnas, en negritas, cursiva, subrayado, etc.). Este esquema recibe el nombre de **XSLT Stylesheet**.

[28] Como se podrá ver en esta y en otras imágenes, se ha suprimido información que no se puede exponer debido a la confidencialidad.

4. CONSTRUCCIÓN DEL CORPUS

A lo largo de este trabajo se ha hecho una serie de exposiciones sobre algunos temas lingüísticos relacionados con la computación, con el objetivo de mostrar de qué forma la tecnología ha revolucionado los métodos de investigación permitiendo obtener resultados de forma más rápida y precisa. Así, es esperado que la construcción del corpus de esta tesis se sirva de dicha tecnología, por lo que los textos que se han recopilado seguirán una serie de procesos que se describirán en este capítulo y que van desde la digitalización hasta el etiquetado del corpus. Cada uno de estos procesos implicará el uso de distintos formatos electrónicos, tales como JPG, TXT y XML.

Como se ha mencionado el objeto de estudio son 121 textos escritos a mano por niños de segundo año de primaria, es decir, niños entre 7 y 8 años de edad. Pero a pesar de que sólo daremos ejemplos de estos 121 textos, es muy importante mencionar que la construcción del corpus del proyecto *Aprender mientras se enseña: una experiencia de acompañamiento en la enseñanza de la Lengua escrita* se hizo con base en los 300 textos que se repartieron entre los lingüistas. Así que es evidente que todos los acuerdos, convenios, decisiones, etc. fueron resueltos por el equipo que se encargó del análisis lingüístico, que como ya se ha mencionado está compuesto por la Dra. Celia Díaz, la Dra. Celia Zamudio, el Dr. Gerardo Sierra, el Dr. Alfonso Medina, el Mtro. Carlos Méndez., la Lic. Argentina Robles, la Lic. Mercedes Tapia, Maria Pucci y su servidora Sandra Richer.

4.1. Digitalización

La digitalización es el proceso de obtener imágenes electrónicas de cualquier formato impreso en material plano como papel, cartón, tela, etc. ya sean con imágenes o textos. Este proceso se lleva a cabo con ayuda de una máquina llamada escáner.

La razón por la cual se acordó digitalizar los textos fue para poder manipular las imágenes de éstos de distintas formas cuantas veces fuera necesario con la seguridad de que los archivos originales no sufrirán alteraciones, mutilaciones, etc.

Asimismo, se acordó guardar algunas copias de las imágenes en diferentes computadoras o dispositivos de almacenamiento masivo[29] para evitar pérdidas y extravíos.

Pero de entre todo esto, la razón más importante por la que decidimos digitalizar los textos fue para compartirlos con todos los investigadores, maestros o alumnos que quisieran consultarlos. Si bien teniendo las imágenes podemos mandar los archivos a quien lo necesite, sabemos que de esta forma el material no llegará muy lejos, así, se llegó a la conclusión de crear una página en internet donde cualquier usuario podrá consultar el corpus y los resultados que arrojará el etiquetado.

4.2. Transliteración

Como ya se ha explicado anteriormente, la transliteración o transcripción de textos consiste en copiar manualmente un texto impreso a algún formato electrónico de escritura. Para este proyecto la transliteración se hizo a formato plano *txt* por dos propósitos:

El primero fue tener los textos de una manera más legible y comprensible, ya que al ser textos escritos a mano por niños que apenas están aprendiendo a escribir muchas veces no es fácil entender lo que dicen los textos, ya sea porque hay muchos tachones, porque las letras están encimadas, porque las grafías de las letras no son claras, etc.

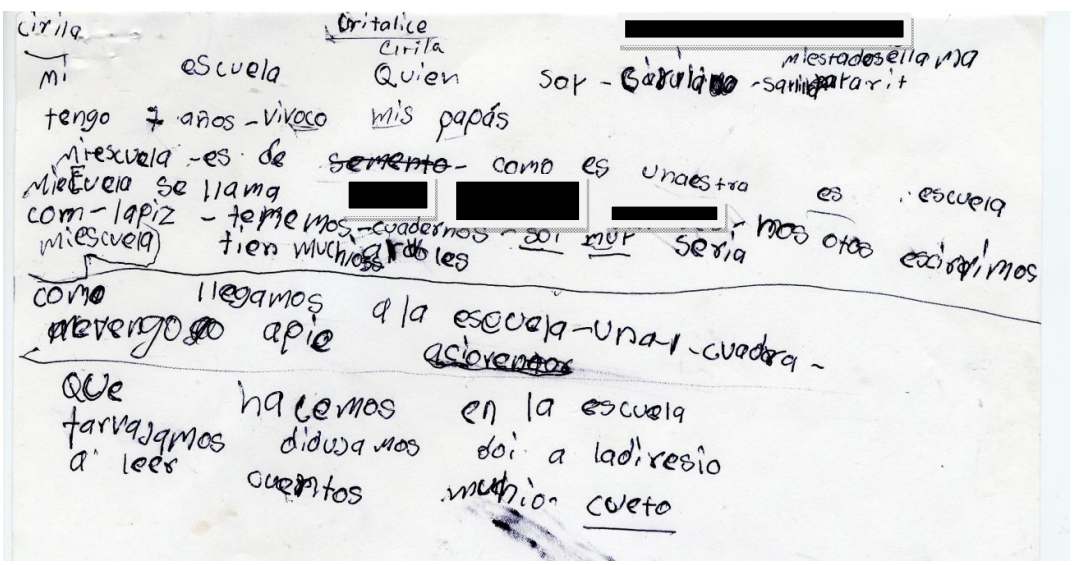


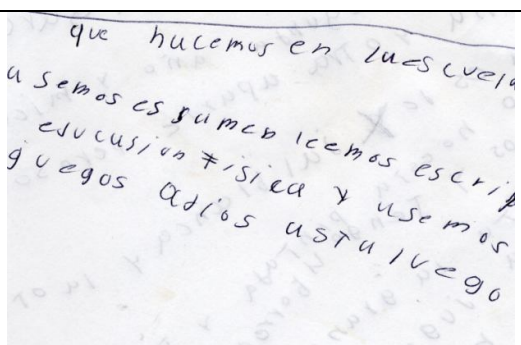
Ilustración G Texto poco legible.

[29] Memorias USB, Ipod, etc.

En varias ocasiones se tuvo que discutir qué era lo que el niño había escrito. Así que después de varios análisis paleográficos, tanto individuales como grupales, obtuvimos textos fáciles de leer sin que éstos dejaran de ser transcripciones fieles a los textos originales, ya que estas transliteraciones a formato txt están sujetas a los siguientes acuerdos:

- Se transliteraron los renglones de la misma forma en como los hizo el niño, es decir, donde el niño termina el renglón la transliteración también lo hace, y se enumeraron cada uno de los renglones, todo esto para tener un mejor control de la ubicación de todos los fenómenos. Algunos niños enumeraron los renglones, así que estos textos tienen doble numeración, la que puso el niño y la que aquí se designó.

- Se respetaron los espacios que dejaban los niños entre letras, palabras, renglones, párrafos, etc., es decir, si el niño pegaba las palabras, separaba una palabra de forma no convencional, dejaba más espacio entre párrafos que entre renglones, etc., la transliteración debía simular los mismos espacios que hacia el niño en su texto, ejemplo:

<table>
<tr>
<td>  </td>
<td>que hacemos en laes cuela

a semos es sumar leemos escribimos
educasion Fisica y asemos
juegos adiós asta luego</td>
</tr>
</table>

Ilustración H Transliteración de los espacios

Esto se hizo debido a que los espacios que hay en los textos nos hablan mucho del concepto de escritura que tienen los niños, por ejemplo, los niños que separan los párrafos con espacios más grandes nos reflejan que tienen una idea más evolucionada de

la escritura, debido a que no sólo están conscientes de que la escritura es una forma de comunicación, sino que además están conscientes de que la escritura tiene sus propias reglas.

Asimismo, los espacios que hay entre las palabras también nos habla de la seguridad que tienen los niños al segmentarlas, por ejemplo, cuando los niños están plenamente seguros de cuando empieza y termina una palabra dejan espacios notables alrededor de ésta, pero cuando tienen duda dejan un espacio muy pequeño, lo cual hace dudar al etiquetador si hay o no espacio ahí.

De esta forma, definir los espacios en los textos txt no fue tarea sencilla, por lo que fue necesario que se hicieran análisis de manera grupal para llegar a acuerdos que ayudaran a una transliteración fiel.

- Se respetaron las autocorrecciones, por ejemplo cuando el niño tacha, encima, repite, etc.

Desde luego en el texto plano no podemos reproducir la tachadura o la forma en cómo se enciman las palabras, así que para marcar estos fenómenos se transliteró la forma "corregida"[30] entre *pico paréntesis* y en la parte final del texto se abrió un espacio para comentarios donde se explica qué es lo que pasa.

La razón por la que se transliteraron estas autocorrecciones fue debido a que son evidencia de que el niño reflexiona sobre lo que escribe, ya que en algunas ocasiones se encuentran casos donde el niño tacha un verbo y lo sustituye por otro, o tacha una frase y la sustituye por otra, ahí el niño está reflexionando sobre lo que escribe, está viendo qué es mejor contar y cuáles son las palabras más adecuadas para ello. Igualmente el niño reflexiona sobre la ortografía de las palabras, duda y analiza la forma correcta, de ahí que encontremos muchos ejemplos de niños que enciman una letra sobre otra, que tachen una "s" final porque se dieron cuenta de la concordancia de número entre palabras, que encontramos una letra chiquita en medio de una palabra porque posiblemente cuando el niño terminó de escribir la palabra se dio cuenta que le falto una letra, etc.

[30] El niño corrige lo que escribe con la o las grafías que cree que son las correctas, pero algunas veces no lo son, incluso llegan haber niños que corrigen algo que estaba bien, y su autocorrección está mal.

- De la misma forma hay otros fenómenos que no se pudieron transliterar, como letras cursivas, letras adornadas, títulos agrandados, subrayados, garabatos, etc. Por lo que al transliterar esto se puso entre *corchetes* y se explicó en el espacio de comentarios cómo estaban construidas estas letras, palabras, frases, oraciones, renglones o párrafos.

La transliteración de esto se hizo debido a que muchas de estas marcas también hablan de una noción de redacción, ya que algunas veces estas marcas resaltan algo importante como un título, una firma, cambio de párrafo, etc.

- La puntuación y acentos también se transliteraron tal como el niño los utilizó, haya sido o no de forma correcta, es decir, si el niño puso puntos sin usar mayúsculas después, usó puntos o guiones para separar las palabras de todo su texto, acentuó de forma incorrecta o no acentuó una palabra que lo necesitaba, etc. Todo eso se transliteró tal cual, sin corregir, agregar o quitar nada.

Esto se hizo debido a que la puntuación y los acentos pueden darle significado al texto dependiendo de cómo se utilicen, de esta forma se quiso registrar el uso que le dan los niños a estas herramientas para conservar el significado que le dieron a sus textos, sea correcto o no a lo que ellos quisieron realmente decir.

En el uso de la puntuación y acentuación también se encontraron casos imposibles de transliterar como acentos sobre consonantes, signos de interrogación invertidos, varios pares de comillas alrededor de una letra, etc. Algunos casos fueron encerrados entre corchetes y explicados en comentarios, otros casos fueron etiquetados como se verá más adelante.

Con respecto a las “i” hubo una serie de inconvenientes debido a que no se llegaba al acuerdo sobre si algunas “i” fueron o no acentuadas por los niños, ya que no se identificaba concretamente qué había puesto el niño sobre la “i”, así que se decidió que si el acento sobre la “i” no era totalmente legible se transliteraría la “i” sin acento para evitar malas interpretaciones, aunque el hecho de transcribir una “i” sin acento cuando el niño sí la puso (ambiguamente) implica una mala interpretación.

Y por último, el segundo propósito por el que se transliteraron los formatos a txt, fue debido a que el preprocesador, el cual es un programa que nos ayudó a etiquetar

algunas cosas de manera automática, sólo acepta este formato para trabajar con los textos.

4.3. Normalización

Una vez que se tuvieron los formatos transliterados notamos que aún había palabras que seguían siendo incomprensibles a causa de omisiones, rotaciones[31], sustituciones, etc. Algunas veces estas palabras llegaban a tener significado gracias al contexto, otras veces sólo eran comprensibles para quién transliteraba debido a que esta persona había encontrado un patrón en las grafías de todo el texto y podía comprender qué era lo que había querido decir el niño, pero otras veces el significado de la palabra se tenía que encontrar con ayuda de los colegas.

Por estas razones se convino que los formatos transliterados no eran suficientes, así que con base en éstos se hicieron las versiones normalizadas, es decir, las versiones corregidas que conciernen a las normas vigentes de la Real Academia Española. Así, estas versiones se realizaron también en formato txt bajo los siguientes parámetros:

- Se respetó los cortes de renglón y la numeración que se hicieron en la transliteración para poder ubicar rápidamente las palabras incomprensibles y comparar las transliteraciones con las normalizaciones fácilmente.

- Se corrigieron:
    - Las faltas de ortografía o sustituciones como *Estevan* por *Esteban*, *miguEL* por *Miguel*.
    - Las omisiones como *escula* por *escuela* o *Maria* por *María*.
    - Elementos agregados como *arrdilla* por *ardilla* o *Marío* por *Mario*.
    - Las rotaciones como *parpue* por *parque*.
    - Las permutaciones como *bontia* por *bonita* o *máma* por *mamá*.

[31] Una rotación es cuando el niño gira una letra, por ejemplo: “d” en lugar de “b”.

Pero no sólo se corrigieron los errores evidentes en la forma física de la palabra como en los ejemplos anteriores, sino también se corrigieron los errores que se hacían evidentes gracias al contexto, como por ejemplo: “hay que a ser” por “hay que hacer”, “la casas blanca” por “la casa blanca”, “dedo hacer” por “debo hacer”, etc.

- Se respetaron los signos de puntuación como se hizo en la transliteración debido a que si se hubiera hecho una “corrección” de éstos cabía la posibilidad de que se cambiara el significado del texto.

- Se corrigieron los acentos quitando los que estaban de más, moviendo los que estaban en la sílaba correcta pero sobre alguna consonante y agregando los que hacían falta.

Por supuesto estas correcciones se hicieron sobre palabras que llevan obligatoriamente acento como “árbol”, “mamá”, “exámenes”, etc. Y sobre palabras que con ayuda del contexto exigían el significado que sólo el acento les puede dar, como “papá”, “habló”, “jugó”, etc.

En raras ocasiones aparecieron palabras ambiguas donde ni el contexto nos pudo ayudar a aclarar el verdadero significado, en estos casos se conservaron las formas transliteradas de las palabras.

- Con respecto a las mayúsculas:
  - Las mayúsculas ubicadas en medio o a final de palabra se sustituyeron por minúsculas, por ejemplo: miguEL -> Miguel.
  - Se corrigieron varios sustantivos propios para que comenzaran con mayúscula.
  - Respetando la estructura que le dio el niño a su texto se sustituyeron las minúsculas por mayúsculas iniciales en las palabras que principiaban párrafos.
  - Con base en la puntuación que hizo el niño se corrigieron las minúsculas por mayúsculas iniciales cuando iban precedidas por un punto.

- En los casos donde el niño usa los puntos para separar las palabras de todo su texto o de algunos párrafos no se sustituyeron las minúsculas por mayúsculas.

La normalización no sólo permitió que el corpus fuera más legible, sino que además nos ayudó a visualizar y terminar de definir el repertorio de todos los fenómenos que nos interesaba analizar, ya que al realizar la transliteración y la normalización notamos muchos otros fenómenos que no habíamos visto en la primera lectura del corpus, de hecho, tanto algunos acuerdos que ya se explicaron como algunos otros que se explicarán más adelante fueron convenidos en la marcha del trabajo.

4.4. Etiquetado

Como se ha visto, constituir un corpus electrónico implica la realización manual de algunas actividades, pero una vez teniendo los primeros formatos electrónicos del corpus podemos servirnos de algunas herramientas computacionales para acelerar y facilitar el resto del trabajo.

Como ya bien se sabe, para analizar y marcar los fenómenos de interés de esta tesis nos servimos del lenguaje de etiquetado XML, pero antes de comenzar a etiquetar de manera manual utilizamos un programa que adaptó el Mtro. Carlos Francisco Méndez. Este programa, llamado *Preprocesador de archivos para XML*, extrae los archivos transliterados, identifica varios elementos y los etiqueta automáticamente. Estos elementos son la numeración, las palabras, la puntuación, los espacios entre palabras y párrafos, los finales de renglón y los comentarios. Asimismo, el Preprocesador coloca al principio del texto una ficha vacía con sus respectivas etiquetas, entre las cuales se pondrá de manera manual información sobre los niños, como se verá en el siguiente punto.

Una vez que el Preprocesador empieza a trabajar con los textos, automáticamente éstos se guardan en formato XML. Así, después de que se obtienen estas primeras versiones etiquetadas de los textos se continúa con el trabajo de manera manual, debido a que sólo un lingüista puede identificar qué fenómenos se presentan y en consecuencia puede poner las etiquetas adecuadas donde corresponde.

A continuación veremos la clasificación tanto de las etiquetas automáticas como de las manuales de acuerdo con su utilidad dentro de siete grupos: las etiquetas que

pertenecen al *encabezado*, las que se relacionan con las *palabras*, las que nos ayudan a etiquetar los *signos de puntuación*, las que nos permiten identificar el inicio y final de las *líneas*, las que identifican la *firma*, con las que señalamos si los niños hicieron *dibujos* en sus textos y finalmente las que nos señalan los *comentarios* del etiquetador.

4.4.1. Encabezado

Como se ha explicado en capítulos anteriores, el ambiente social en el que se desenvuelve el niño influye en gran manera en su capacidad de aprendizaje y desarrollo académico. Está comprobado que los niños que se desenvuelven en ambientes propicios para su desarrollo y aprendizaje logran tener un conocimiento de lo que es la escritura desde antes de que aprendan a escribir, y esto les ayuda de manera considerable a adquirir fácilmente la lengua escrita, a diferencia de los niños que se desenvuelven en zonas donde hay un bajo nivel de educación y poco interés por el discurso escrito.

Por esta razón, cuando se recoge el corpus es muy importante obtener y registrar información socioeconómica para analizar cómo se adquiere el discurso escrito en los distintos rubros sociales, explorar cuáles pueden ser las posibles causas que obstaculizan o estimulan el aprendizaje, examinar si la facilidad de aprendizaje varía entre géneros, etc. Todo esto con el fin de generar nuevas metodologías didácticas o corregir las existentes y ayudar a que los niños desarrollen competencias de escritura y lectura, las cuales les ayudarán a lo largo de su vida a comprender mejor los textos que se les presenten y en consecuencia a adquirir nuevos conocimientos.

Los estudios lingüísticos consideran, cada vez con mayor énfasis, las relaciones entre lengua y sociedad implicadas en los análisis sobre la lengua en uso, y aunque el proyecto *Aprender mientras se enseña: una experiencia de acompañamiento en la enseñanza de la Lengua escrita* no pretende hacer ese tipo de análisis, se consideró de vital importancia exponer la información social de cada uno de los textos, ya que uno de los objetivos es que el corpus sirva como material para diversas investigaciones en distintas áreas. De esta forma, además de que esta información ayudará a darle veracidad y validez al corpus, también ayudará a que otros investigadores complementen su análisis lingüístico con la razón social que les ofrecemos.

Así, teniendo en cuenta la importancia de los datos sociales, se resolvió incluir la información de cada niño al principio de sus textos dentro del etiquetado, esto permitió

tener la información totalmente a la mano, evitando por supuesto que hubiera intercambios, confusiones o extravíos de datos.

Dentro del etiquetado, la información se ordenó en la ficha vacía que integró automáticamente el preprocesador.

A continuación tenemos como ejemplo una ficha o encabezado donde se señala en negritas o con tres asteriscos[32] toda la información que se puso manualmente:

<encabezado>
<niño nombre="*******" sexo="**F**" edad="*******"
fechanacimiento="*******" grado="*******" grupo="*******"/>
<maestro nombre="*******"/>
<director nombre="*******"/>
<escuela nombre="*******" clave="**01**" zona="**04**" sector="**10**"/>
<acompañante nombre="*******"/>
<archivo nombre="**N041001EscF10.txt**"/>
<imagen archivo="**N041001EscF10.jpg**"/>
</encabezado>

En sí, la información que se solicita es clara, pero hay que especificar algunos puntos:

1. Dentro de la información que se solicita sobre la escuela se pide la "clave". Si a un etiquetador le tocaron tres escuelas de un mismo sector y una misma zona, el etiquetador las enumera para diferenciarlas y ese número es la clave.
2. El **acompañante nombre** es el maestro acompañante, el que capacitó a los maestros y recolectó el corpus, tanto en los grupos donde capacitó al maestro como en los grupos donde no lo hizo.
3. El **archivo nombre** es el nombre que se le asignó a las imágenes y a las transliteraciones, y se decodifica de la siguiente forma:
    - La primera letra designa si el texto proviene de un grupo donde el maestro **S**í participó en las capacitaciones, o **N**o lo hizo.
    - Los siguientes dos números son los del sector de la escuela.
    - Los dos números que siguen corresponden a la zona.
    - El número de la escuela se señala con los dos números que continúan, es la "clave".
    - La marca Esc señala que se va a trabajar con el texto en prosa. Se hizo además una recolección de poemas escritos por los niños, pero ese material aún no se analiza.
    - La siguiente letra señala el sexo del niño: **M**asculino o **F**emenino.
    - Y por último se marca el número del niño, el cuál designo el etiquetador.

Algunos niños escribieron más de una hoja, así, como se podrá ver en ejemplos más adelante, algunos nombres de archivo tienen una letra al final, ésta marca el número de hoja, por ejemplo, si el nombre de un texto termina con una "d" significa que esa hoja

[32] Los asteriscos sustituyen información que no se puede exponer debido a la confidencialidad.

es la 4ª que escribió el niño. Cuando no hay ninguna letra al final del nombre del archivo, como en el ejemplo de arriba, significa que el niño sólo escribió una hoja.

4.4.2. Palabras

Tomando en cuenta el objetivo de esta tesis es de vital importancia etiquetar las palabras, y las etiquetas a las que recurrimos para ello fueron **<g></g>**[33]. Estas etiquetas son asignadas por el preprocesador, el cual se encargó de identificar cada una de las palabras basándose en el criterio de que cada una de ellas son conjuntos de letras separados por espacios. Sin embargo conviene hacer algunas precisiones, debido a que en muchas ocasiones etiquetó como una palabra estructuras como "miescuelaestabonita" o como dos palabras estructuras como "mies cuela", por lo que se tuvieron que hacer los ajustes de manera manual, pero éstos se explicarán más adelante cuando describamos las demás etiquetas, las cuales por supuesto se colocaron dentro de las etiquetas <g><g/> para ayudarnos a corregir o complementar el etiquetado del preprocesador.

Hay que recordar que los textos que se etiquetaron fueron los que se transliteraron, pero como ya sabemos, estos últimos textos aún tenían problemas para ser totalmente legibles, de ahí que hayamos decidido hacer las versiones normalizadas, pero dentro del etiquetado hacer una versión normalizada nos pareció exceso de material y de trabajo, por lo que optamos por recurrir a una etiqueta que engloba los beneficios de la normalización, el atributo **<n="">**. Así, entre las comillas de este atributo se escribieron manualmente las normalizaciones de las palabras que tenían fenómenos ortográficos, por ejemplo, si el niño escribió "escula" el atributo se usa de esta forma:

<g **n="escuela"**>escula</g>

Este atributo sólo se usa cuando el niño no escribe la palabra de acuerdo con la normalización ortográfica que utilizamos, pero si el niño escribió bien la palabra no se usa el atributo n=""; por ejemplo, si el niño escribió "miescuela" o "laca saes" no se usa

[33] Esta es la misma etiqueta con la que se señalan las palabras en el CHEM.

el atributo n=“” porque aunque ocurren fenómenos de segmentación, las palabras están escritas ortográficamente correctas.

El atributo n=“” es importante, ya que al tener la función de normalizar, tiene uno de los beneficios más importante de ésta, la desambiguación. Esta etiqueta nos ayuda a identificar el verdadero significado de las palabras, el cual sólo se encuentra gracias al contexto o al patrón de grafías que encuentra el etiquetador, por ejemplo:

1. … después de comer <g n=“debo”>dedo</g> hacer la tarea …
2. … se <g n=“rió”>vio</g> con los niños en el patio …

Con la palabra “dedo” nadie supondría sin ayuda del contexto que hubo una rotación de la “b”, y que en lugar de “dedo” lo que realmente quiso decir el niño fue “debo”. De la misma forma en el ejemplo 2 muchos deducirían al ver la palabra “vio” que es una conjugación del verbo “ver” como sinónimo del verbo “citar”, pero en realidad el etiquetador descubrió que muchas de las “r” que escribió el niño las escribió como “v”, y por consiguiente descubre que “vio” significa en realidad “rió”.

4.4.2.1. Segmentos

Dentro del discurso escrito, más que simples espacios, la correcta segmentación significa la concretización de la forma física de cada palabra y el claro entendimiento del significado de cada una de ellas. Al etiquetar los espacios que utilizan los niños pretendemos simplemente conservar la segmentación para después analizar los fenómenos relacionados con ella. De esta forma creamos la etiqueta **<e/>** la cual sustituye los espacios que hay entre palabras tal y como lo hizo el niño, así, si encontramos casos como *mies cuela esta bonita* el preprocesador lo etiquetará:

<g>mies</g>**<e/>**
<g>cuela</g>**<e/>**
<g>esta</g>**<e/>**
<g>bonita</g><[34]

[34] Esta etiqueta se dejó vacía debido a que puede haber un espacio o un cambio de línea, es decir que hay dos posibilidades de etiqueta.

En este ejemplo podemos ver que hay fenómenos que saltan a la vista, los cuales son conocidos como hipersegmentación e hiposegmentación. El primer fenómeno señala que una palabra fue segmentada en dos o más segmentos y la hiposegmentación nos indica que dos o más palabras fueron pegadas o juntadas como si constituyeran una sola. Al etiquetar estos fenómenos veremos qué tan frecuentemente caen los niños en ellos y qué porcentaje ocupan en los textos, esto nos dará un panorama de qué tan complicado es aprender a segmentar correctamente.

Cuando en el texto se encuentra una hipersegmentación como “es cuela” el preprocesador asigna un par de etiquetas <g> a cada uno de los segmentos:

<g>es**</g>**<e/>
<g>cuela**</g>**<

Al corregirlo se quitaron algunas etiquetas <g> y se agregó la etiqueta **<hiperseg/>** quedando de la siguiente forma:

<g>es**<hiperseg/>**<e/>cuela**</g>**<e/>

Este formato le permite al programa contar la palabra como una sola, permitiéndonos ver a la vez el fenómeno de hipersegmentación, ya que agregamos la etiqueta <hiperseg/> y además no quitamos la etiqueta de espacio <e/> entre los elementos.

Asimismo, si el niño separó una palabra poniendo los segmentos en renglones diferentes, se corrige de forma similar. Ejemplo:

<g>escuela</g><e/>
<g>estud**</g>**<l/>
<nl num=“7”/>**<g>**iamos**</g>**<e/>

<g>escuela</g><e/>
<g>estud**<hiperseg/>**<l/>
<nl num=“7”/>iamos**</g>**<e/>

A pesar de que hay dos etiquetas más entre las etiquetas <g></g> el programa siempre reconocerá como una sola palabra el conjunto de letras que esté dentro de las etiquetas <g><g/> aunque haya muchas otras etiquetas dentro de ellas. Esto nos permitirá marcar

los diferentes fenómenos que una sola palabra puede presentar, como se verá más adelante.

En los casos donde el niño separó una palabra poniendo los segmentos en renglones diferentes y usó un guión al final del primer renglón, no se considera hipersegmentación aunque haya segmentado la palabra de forma silábicamente incorrecta, esto se convino debido a que el hecho de que pusiera el guión significa que el niño sabe que esa palabra no se separa, y que si lo hizo fue simplemente porque no le cupo en el renglón, por ejemplo:

… esta cer- ca de ...	... a la esc- uela caminando ...
<g>esta</g><e/>	<g>a</g><e/>
<g>cer<r c="Fg"/>---</r><l tipo="guion">	<g>la</g><e/>
<nl num="5"/>ca**</g>**<e/>	**<g>**esc<r c="Fg"/>---</r><l tipo="guion">
<g>de</g><	<nl num="10"/>uela**</g>**<e/>
	<g>caminando</g><

Pero aunque no usemos la etiqueta de hipersegmentación, nótese que se corrigen las etiquetas <g></g> para que se consideren los segmentos como una sola palabra.

En contraparte, cuando encontramos en un texto ejemplos como "micasa", el preprocesador lo etiqueta así:

<g>micasa</g><e/>

Para corregirlo se agregaron etiquetas <g> y se uso la etiqueta **<hiposeg/>** que indica la hiposegmentación que hubo, quedando:

<g>mi**<hiposeg/>**</g>
<g>casa</g>

Así el programa podrá contar ambas palabras mientras nos permite ver el fenómeno de hiposegmentación, ya que a pesar de que las palabras están en diferentes renglones para una mejor lectura del etiquetado, en la hoja de estilo[35] aparecerán como las escribió el niño debido a que no pusimos la etiqueta de espacio <e/> entre ellas.

[35] La hoja de estilo es el resultado del etiquetado, es decir, es la presentación final del corpus y lo describiremos más adelante.

Como hemos visto, pueden aparecer muchos fenómenos dentro de una misma palabra, pero lo más “complicado” fue cuando aparecían los dos fenómenos de segmentación muy próximos, ya que a diferencia de los otros fenómenos no basta con sólo poner las etiquetas correspondientes, sino que se tiene que agregar o quitar etiquetas de palabra <g> para que el programa pueda contar de manera correcta el número de palabras y además se tiene que arreglar la presentación para facilitar la lectura del etiquetado; entonces el ejemplo:

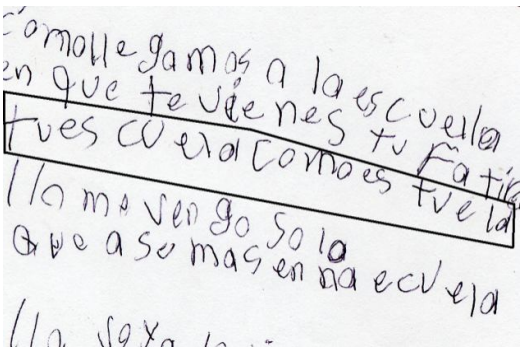


Ilustración I. Hipo e hipersegmentaciones.

tues cu elacomoes tue la

Es etiquetado por el preprocesador de la siguiente forma:

<nl num=“15”/><g>tues</g><e/>
<g>cu</g><e/>
<g>elacomoes</g><e/>
<g>tue</g><e/>
<g>la</g><l/>

La dificultad para corregir este etiquetado radica en que hay que estar atentos a las etiquetas que se agregan y se quitan, porque de ello depende un correcto conteo de los fenómenos para poder hacer un análisis del porcentaje de ocurrencia de éstos, por tanto, el etiquetado correcto sería:

<nl num="15"/><g>tu**<hiposeg/>**</g>
<g>es**<hiperseg/>**<e/>cu**<hiperseg/>**<e/>ela**<hiposeg/>**</g>
<g n="cómo">como<omi tipo="acento"/>**<hiposeg/>**</g>
<g>es</g><e/>
<g>tu**<hiposeg/>**</g>
<g n="escuela">e<omi tipo=“scue”/>**<hiperseg/>**<e/>la</g><l/>

También es importante mencionar que las etiquetas se pusieron tantas veces se presentaron los fenómenos, aún cuando se repitieron las mismas etiquetas dentro de una misma palabra, ya que esto permite también llevar un conteo adecuado de los fenómenos.

4.4.2.2. Fenómenos ortográficos

A lo largo de esta tesis hemos estado haciendo énfasis en que el discurso escrito no consiste meramente en escribir un montón de palabras, sino que el discurso escrito es un medio de comunicación tan importante como el discurso oral y tiene una estructura específica para lograr una clara transmisión de información. Como parte de esa estructura, además de una correcta segmentación de palabras es necesario saber utilizar las grafías de escritura (las letras) correctamente, ya que de lo contrario puede haber ambigüedades.

Las primeras grafías que nos llamó mucho la atención etiquetar fueron las mayúsculas, ya que además de ser parte de una buena redacción, muchas veces escribir una palabra con mayúscula le da un significado diferente a las palabras, por ejemplo, no es lo mismo "abril" que "Abril" ya que la primera palabra se refiere a un mes y la segunda a un nombre propio, o las frases nominales "el señor" y "El Señor" donde la primera se refiere a un señor cualquiera y la segunda frase se refiere a Dios.

Pero además de desambiguar, el uso de mayúsculas nos ayuda a identificar los nombres propios rápidamente, así como inicios de párrafo, o el final de un enunciado y el comienzo de otro en el mismo renglón. Claro está que estas últimas marcas también implican el uso de puntos, pero de eso hablaremos más adelante.

Habiendo explicado la importancia de las mayúsculas, es evidente que nos interesa mucho la manera en cómo las utilizan los niños, por lo que creamos dos etiquetas para identificarlas: La primera es **<mayIni/>** y la utilizamos para señalar las mayúsculas que fueron escritas en *el lugar apropiado*, por ejemplo:

<g>María**<mayIni/>**</g>

¿A qué nos referimos con *el lugar apropiado*? Pues bien, cuando se analizó el corpus al momento de transliterar y normalizar nos dimos cuenta que había gran cantidad de ejemplos donde los niños escribieron mayúsculas a lo largo de la palabra, lo cual no es *el lugar apropiado*, ya que éste es únicamente a principio de palabra. Y algo muy importante es que en español por cada palabra que lo requiera sólo se usa **una** mayúscula, con excepción de las palabras que representan siglas.

De esta forma la segunda etiqueta que utilizamos para marcar mayúsculas es **<mayInter/>**, con la cual señalamos todas esas mayúsculas que no están donde corresponden, ejemplo:

<g n="escuela" >esCueLa**<mayInter/>**<sus tipo="cxC"/>**<mayInter/>**<sus tipo="lxL"/> </g>.

El objetivo de utilizar esta etiqueta es identificar qué tan frecuentemente los niños comenten el error de escribir una mayúscula en vez de una minúscula y ver si hay un patrón de grafías recurrente. Lo cual puede ayudar a otras investigaciones a determinar cuál es el problema que orilla a los niños a hacer esto, y de ello elaborar una solución.

En relación con esto encontramos que algunos niños entienden el uso de las mayúsculas y la importancia que representan, pero pareciera que por alguna razón no han logrado aprender u olvidan la forma gráfica de algunas mayúsculas, a lo que resuelven, curiosamente, en agrandar las minúsculas de la letra que requieren para darle a su texto la apariencia más correcta, por ejemplo:

<g n="jugamos"> <j>ugamos**<agrandaMin/>**</g>[36]

Ilustración J Agrandamiento de minúscula

Nos pareció importante señalar este fenómeno porque no se trata si el niño sabe o no cuándo y cómo se usa una mayúscula, se trata de que el niño está consciente de esas reglas y trata de solucionar de una manera, en lo personal, muy inteligente el inconveniente de no recordar la grafía.

Como se puede ver, las etiquetas nos ayudan a hacer un conteo de los fenómenos que se presentan en los textos, pero algunas veces para complementar la información que nos dan algunas etiquetas se necesitan otras etiquetas, y este es el caso de la etiqueta **<sus/>**, la cual nos ayuda con la etiqueta <mayInter/> a identificar qué grafías fueron las que se intercambiaron. Esto nos permitirá al final hacer un conteo de éstas y ver si realmente hay un patrón de grafías recurrentes en la alternancia de minúsculas por

[36] Los símbolos < y > son equivalentes al pico paréntesis que se uso en la transliteración para marcar los fenómenos que no se podían transliterar, dentro del etiquetado XML los pico paréntesis sólo se pueden usar para las etiquetas.

mayúsculas. Pero aparte de esta función, la etiqueta <sus/> se encarga de identificar todos los intercambios de grafía que hubo en el texto, ejemplo:

<g n="baño"/>ballo**<sus tipo="ñxll"/>**</g>
<g n="zapatos"/>sapatos**<sus tipo="zxs"/>**</g>

Con esta etiqueta identificamos todas las faltas de ortografía comunes como la sustitución de *v* por *b* o de *z* por *s*, etc. Las cuales podrían ser causadas por la semejanza fonológica que tienen en el español de México.

De la misma forma encontramos sustituciones entre grafías que pertenecen a fonemas diferentes pero que son muy próximos en el punto de articulación, como la *ñ* y la *ll*, o la *r* y la *l*, entre otras. Esto nos permite ver que realmente en la oralidad hay líneas muy delgadas entre unos y otros fonemas, por lo que es común que en el proceso de adquisición de la escritura los niños tengan dudas sobre qué grafía corresponde a tal o cual fonema.

Además de ayudarnos a encontrar problemas ortográficos, la etiqueta <sus/> nos ayudó también a identificar muchos errores a nivel morfológico relacionados con la concordancia de género entre sustantivos, verbos, adjetivos, complementos, etc. Obteniendo de ello ejemplos como: "mi casa es amarillo", "los niñas lloran", "a mi perro la llevo a pasear", etc.

Asimismo, la etiqueta <sus/> nos permitió ver que los niños tienen problemas para diferenciar unas letras de otras cuando gráficamente son muy parecidas, como *h* y *n*; o tienen problemas cuando los movimientos para escribir dos letras son muy similares, provocando que por ejemplo muchas *r* parezcan *v*, algunas *q* parezcan *a*, etc.

Pero en los casos donde se intercambia la grafía *p* por la *q*, o la *b* por *d*, no creemos que haya una sustitución, más bien definimos éste último fenómeno como rotación, debido a que parece que la letra giró o rotó. Por tanto, utilizamos la etiqueta **<rotac tipo=""/>** en cuyo atributo señalamos la letra que se rotó utilizando su forma correcta, por ejemplo:

<g n="dibujamos">didujamos**<rotac tipo="b"/>**</g>

Creer que la escritura es una transliteración de la oralidad nos puede hacer caer en muchos errores ortográficos, ya que como se ha visto la fonología puede provocar

algunas confusiones en grafías, y no sólo en relación con la sustitución entre éstas, sino también en relación con omisiones. Esto se debe a que muchas veces en la oralidad omitimos ciertos elementos que son visibles solamente en la escritura, y el ejemplo más recurrente de esto es la omisión de la *h*. De aquí que hayamos acordado utilizar la etiqueta **<omi/>,** la cual nos permite resaltar todas las omisiones que se presentaron en los textos, por ejemplo:

<g n="Abraham">Abram**<omi tipo="ha"/>**</g>
<g n="televisión">telebisio<sus tipo="vxb"/>**<omi tipo="acento"/><omi tipo="n"/>**</g>

Además de ayudarnos a ver todos los elementos de la escritura que se omiten en la oralidad, la etiqueta <omi/> nos permitió ver que muchas de éstas omisiones también son causadas debido a que los niños posiblemente no logran hacer concordancias de número, por lo que es común encontrar frases como: "me gust<u>a</u> los chocolate<u>s</u>", "hay much<u>o</u> niño<u>s</u>", "lo<u>s</u> juguet<u>e</u>".

Con todo lo anterior podemos resolver que si la oralidad muchas veces no es de gran ayuda para escribir bien las palabras, es necesario memorizar cómo están constituidas, y una prueba más de ello es cuando se usan las letras correctas pero en orden equivocado, lo cual es considerado como una metátesis o *permutación*, por ejemplo: "co<u>sn</u>truyo". Nos interesó etiquetar este fenómeno porque no se trata de adivinar que letra pertenece a qué fonema, sino se trata de ver qué tan difícil es para los niños entender el orden de los fonemas en una palabra. Para ello nos servimos de la etiqueta **<per/>** la cual nos ayuda a reconocer en qué tipo de construcciones silábicas ocurre esto y qué letras son las más comunes de confundir. Pero además de las letras, esta etiqueta nos ayuda también a identificar permutaciones de acentos, es decir, cuando un niño pone un acento en una palabra que sí lo requería pero no lo pone en la vocal correcta, o incluso lo pone sobre alguna consonante[37], por ejemplo:

<g n="árbol">arból**<per tipo="acento"/>**</g>

[37] En estos casos se transcribió el acento después de la consonante, por ejemplo: "mam´a", y además de usar la etiqueta de permutación se usaron corchetes y se agregó un comentario en la parte de abajo para explicar mejor lo que ocurrió.

Así como se encontraron elementos omitidos, también se encontraron elementos de más, es decir agregados, por tanto se decidió crear la etiqueta **<agre/>**. Por supuesto encontramos fenómenos que pueden relacionarse con la oralidad como la vinculación entre las palabras "lejos" y "cercas", pero además encontramos muchos otros fenómenos donde los niños agregaron letras por posibles otras razones que tienen que ver más con la escritura, ya sea porque se dejaron llevar por las palabras cercanas, como en el enunciado "las bancas están forradas de *rositas*", porque vincularon palabras como "María" y "Maríо", porque no supieron hacer concordancias como en "hay una maestra que se llevan a los niños", etc.

4.4.2.3. Autocorrecciones

Muchas de las etiquetas anteriores nos han permitido ver de qué forma los niños resuelven los distintos problemas que se les presentan al momento de escribir, lo cual es importante debido a que con el análisis de esos fenómenos podemos suponer cómo reflexionan los niños sobre la escritura. De la misma forma, las autocorrecciones podrían ser otra forma de reflexión que consideramos de vital importancia señalar en el corpus, ya que con ayuda de la etiqueta **<correccion/>** podemos observar las viables reflexiones que hace el niño sobre las palabras una vez que las ha terminado de escribir, a diferencia de las etiquetas antes mencionadas que nos ayudan a observan las posibles reflexiones que hace el niño mientras está escribiendo las palabras.

Por supuesto, usar solamente la etiqueta <correccion/> no fue suficiente para obtener toda la información que nos interesaba, así que como parte del atributo de ésta nos servimos de dos nombres de los elementos de las etiquetas anteriores, los cuales son: <correccion tipo="**agre**"/>, <correccion tipo="**susti**"/>. Al igual que los elementos, estos atributos se encargan de describirnos cuáles elementos fueron insertados o sustituidos, pero a diferencia de los elementos, los atributos nos señalan la realización de estos fenómenos de manera consciente por parte el niño; por ejemplo, si el niño escribió "pisarrón" usamos la etiqueta <sus tipo="zxs"/>, pero si en vez de dejar la palabra así el niño puso una *z* sobre la *s* usamos la etiqueta <corrección tipo="susti"/>. De esta forma en el primer ejemplo la etiqueta está señalando un error del niño y en el segundo la etiqueta está exponiendo la reflexión del niño.

Además de estas dos maneras en las que se evidencia las reflexiones de los niños sobre la escritura, encontramos otra que consiste en que los niños eliminan una letra,

una palabra, una frase o toda una oración. Algunas veces al eliminar estos elementos, la redacción ya está corregida, pero en otras ocasiones los niños eliminan algo con la finalidad de sustituirlo por otra cosa que exprese mejor lo que tratan de decir.

Es importante mencionar que en los elementos eliminados también podía haber fenómenos ortográficos, pero decidimos no marcarlos debido a que al ser eliminados por el niño ya no se consideran como parte del texto.

En la imagen K veremos algunos ejemplos de estos tres casos de autocorrección.

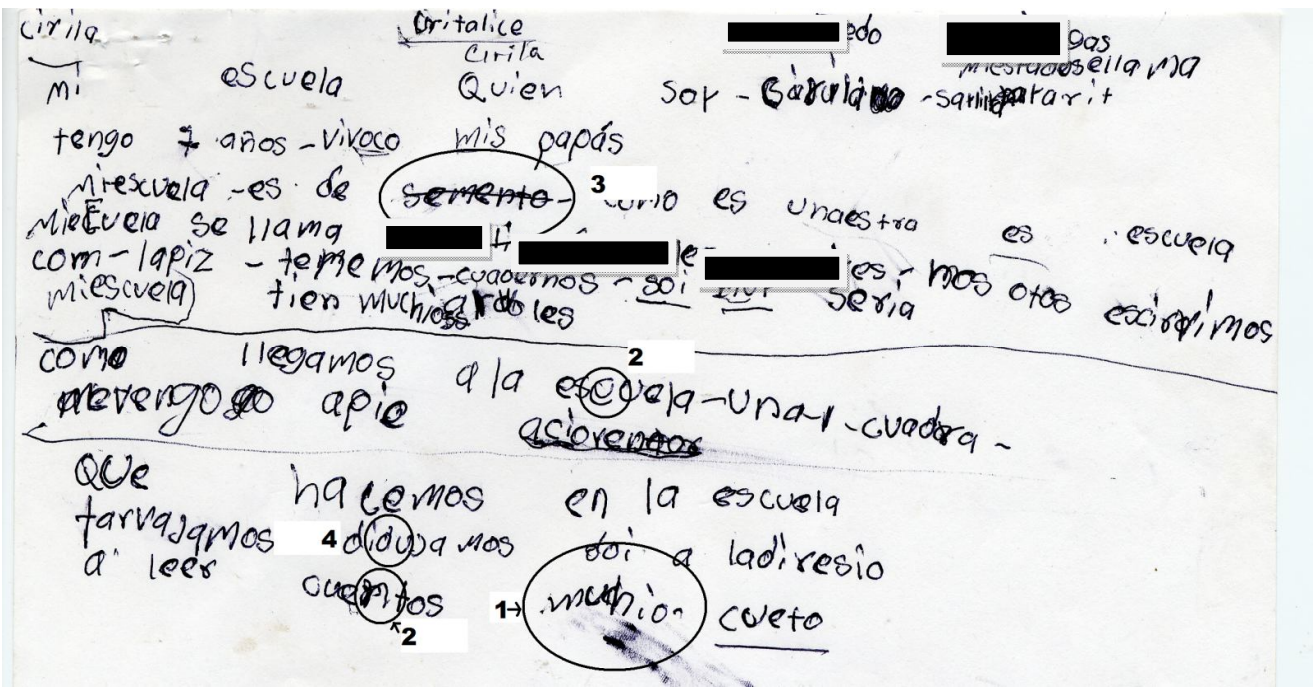


Ilustración K 1. Inclusión. 2. Sustitución. 3. Eliminación.

4.4.3. Signos de puntuación

Como ya se ha mencionado antes, los signos de puntuación son parte fundamental de la estructura del discurso escrito, ya que contribuyen a darle significado y orden a las ideas. Esto lo logran debido a que señalan pausas en el texto y eso ayuda a jerarquizar las ideas principales de las secundarias, además de delimitar frases y párrafos. Asimismo, algunos signos le dan a las frases una tonalidad que también dota de significado.

De esta forma, es muy importante saber algunas reglas básicas de la puntuación para poder expresar lo que realmente se quiere decir, y aunque en nuestro corpus los

niños usan la puntuación básica, nos pareció muy interesante analizar qué signos de puntuación usan y de qué manera.

Dependiendo de la posición y el tipo de puntuación, el texto puede adquirir uno u otro significado, por tanto no nos pareció correcto modificar de ninguna forma la puntuación que utilizan los niños dentro de la normalización, ya que al modificarla corríamos el riesgo de perder el verdadero significado del texto.

La etiqueta de la que nos servimos fue **<r></r>**, a la cual le agregamos dos atributos, uno que se implementó de forma automática y otro que se introdujo manualmente. El primero es el atributo <r **c=“”**></r>, y en él se especifica qué tipo de puntuación es, ya sea:

a) c=“*Fg”* para guión
b) c=“*Fe”* para comillas
c) c=“*Fsp”* para signo de pesos ($)
d) c=“*Ft”* para signo de porcentaje
e) c=“*Fp”* para punto
f) c=“*Fx”* para punto y coma
g) c=“*Fd”* para dos puntos
h) c=“*Fc”* para coma
i) c=“*Fia”* para signo de interrogación
j) c=“*Faa”* para signo de admiración

El segundo atributo es <r c=“Fp” **punto=“”**></r> y como se puede ver éste sólo puede usarse cuando en *c* se está señalando un punto, el cual es especificado en este segundo atributo.

Para especificar el tipo de punto nos basamos únicamente en la ubicación que tenían dentro del texto, por ejemplo, si veíamos un punto en medio de un renglón lo consideramos como punto seguido aunque haya habido o no una palabra con mayúscula inicial posteriormente. Así logramos clasificar cuatro tipos de punto: seguido, aparte, final y el que se usa después de una abreviatura. Obteniendo etiquetas como esta:

<r c=“Fp” punto=“**abrev**”> . </r>

Hay casos muy frecuentes donde el niño usa los guiones o puntos como herramientas para separar las palabras de todo su texto o de algunos párrafos, en estos casos se hizo

una pequeña modificación en el atributo c="", se agregó una **s** que significa "separación", así las etiquetas quedan de la siguiente forma: <r **c="Fgs"**> o <r **c="Fps"**>. Esto se hizo para distinguir las funciones que entonces adquieren los guiones y los puntos, ya que los que se presentan en este último caso no tienen un valor verdaderamente ortográfico, sólo son pequeñas marcas que ayudan al niño a separar las palabras, es decir, tienen valor de segmentación en vez de puntuación. Por tanto, es de suponerse que en los textos donde se usó esta técnica nos servimos de los puntos para encontrar también casos de hipo e hipersegmentación. De esta forma, si encontrábamos ejemplos como ".mi.es.cuela.", marcábamos una hipersegmentación, o una hiposegmentación en los casos como ".años.ytego.cuatro.caballos."

4.4.4. Líneas

Como ya hemos dicho, los puntos no sólo le dan significado al texto, sino también contribuyen a darle una estructura, la cual, como ya hemos mencionado, nos indica que el niño tiene una conceptualización de la escritura más consolidada; de esta forma, como parte también importante para el análisis estructural del texto, convenimos en analizar de igual manera los renglones, ya que en una primera lectura de corpus encontramos fenómenos interesantes.

Primeramente, como simple herramienta consideramos importante conservar en el etiquetado la numeración que hicimos en la transliteración, ya que realmente nos fue de mucha ayuda para identificar rápidamente los fenómenos que describíamos en la sección de comentarios. De esta forma designamos la etiqueta **<nl/>** con el atributo **"num"** donde se designa el número de renglón, por ejemplo <nl num="15"/>.

De la misma forma nos pareció interesante analizar la forma en cómo los niños terminan los renglones cuando una palabra ya no cabe en el espacio que les queda en la hoja. Por tanto, creamos la etiqueta **<l/>** para marcar los finales de renglón y conforme fuimos avanzando en el análisis creamos los atributos correspondientes para esta etiqueta, dando como resultado cuatro tipos de cierre: uno es la forma correcta que consiste en poner la palabra incompleta y usar un guión para así completar la palabra en el siguiente renglón; la otra forma que encontramos es similar pero no se usa el guion; la tercera opción fue que los niños hicieron las últimas letras de la palabra más chiquitas o muy pegadas entre ellas para que cupieran como se ve en la ilustración L; y la cuarta es cuando a pesar de que pegaron las letras o las hicieron más chiquitas, aún no cupo la palabra y se vieron en la necesidad de poner un guion para completar la palabra en el renglón siguiente. De esta forma se crearon los atributos:

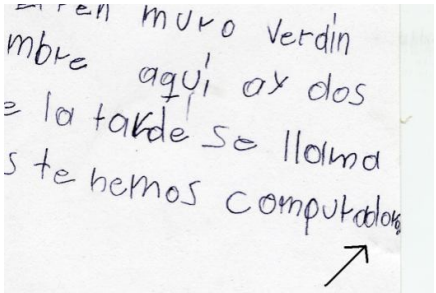


Ilustración L Letras compactadas.

<l tipo="**conguion**"/>

<l tipo="**singuion**"/>

<l tipo="**compac**"/>

<l tipo="**guioncompac**"/>.

4.4.5. Firma

Hasta este punto ya hemos analizado muchos elementos que nos han ayudado a darnos una idea de la forma en que los niños conceptualizan la escritura y qué tanto están conscientes de las reglas que la constituyen. Pero aún con todo lo anterior todavía encontramos elementos muy importantes, uno de ellos es utilizar una firma al final, la cual realmente es exclusiva del discurso escrito, ya que por ejemplo en la oralidad después de que terminas de decir algo no dices al final quién eres, pero en la escritura si es lógico poner tu nombre al final de lo que escribes, incluso en algunos contextos es necesario hacerlo; por tanto, el hecho de que algunos niños pongan su firma al final implica que tienen el concepto de escritura más consolidado, y no sólo eso, sino que además cabe la posibilidad de que estén conscientes de los tipos de escritura que hay, es decir que ponen su firma al final porque saben que están escribiendo una carta, no un cuento, o un resumen. Ya que el hecho de que pongan su nombre no es para

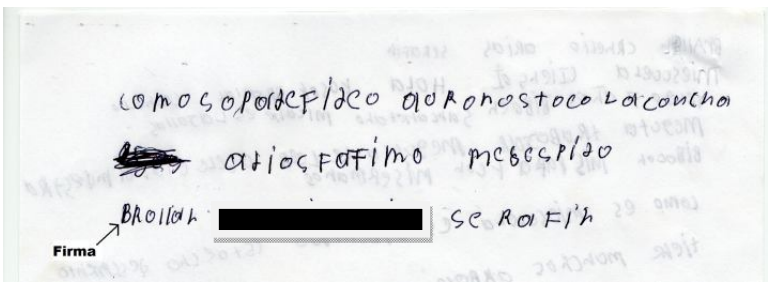
identificarse, debido a que en todos los textos los niños tienen que escribir su nombre completo al principio. Entonces, esa firma es una muestra de que los niños tienen una definición más consolidada de la escritura y de los tipos de texto que se pueden realizar.

Por tanto, para resaltar este elemento que presentan algunos niños nos servimos de la etiqueta **<firma tipo="">**, en cuyo atributo señalamos las características con las que los niños hacían sus firmas:

tipo="a": Señala que el niño simplemente escribió su nombre.
tipo="b": Señala que el niño escribió su nombre con letra diferente a la del resto del texto, por ejemplo más grande, cursiva, subrayada, etc.
tipo="c": Se usa este atributo cuando el niño simplemente hace un garabato.
tipo="d": Nos indica que el niño puso un garabato antes de su nombre.
tipo="e": Nos indica que el niño puso un garabato después de su nombre.
tipo="f": Se utiliza este atributo cuando el niño hace una firma tipo "b" pero agrega un garabato.

Muchas de estas firmas son introducidas por palabras o frases como: "firma:", "hecho por:", "atentamente", etc. Pero la etiqueta se utiliza de la misma forma tanto en firmas introducidas como en firmas que no lo están.

En algunas firmas pueden presentarse todos los fenómenos que van dentro de las etiquetas <g></g>, por lo que las etiquetas <firma></firma> pueden encerrar a las etiquetas <g></g> y éstas a su vez encerrar los fenómenos que presenta la firma, como si fueran fractales, por ejemplo:



Brollan *** se Ra fin

Ilustración M. Firma con fenómenos.

<nl num="22"/>**<firma tipo="a">**<**g** n="Bryan">Brollan<mayNP/><agre tipo="o"/><sus tipo="yxll"/>**</g>**<e/>
<**g**> *** **</g>**<e/>
<**g** n="Serafín">se<sus tipo="Sxs"/><hiperseg/><e/>ra<hiperseg/><SUS tipo="rxR"/><e/>fin<omi tipo="acento"/>**</g></firma>**<l/>

4.4.6. Dibujos

Ilustración N. Dibujo.

Como ya se ha mencionado otras veces, en los textos de los niños suelen presentarse expresiones que no pueden transcribirse, y los dibujos son una de estas expresiones, así que para no omitir el hecho de que el niño hace estos dibujos, utilizamos la etiqueta <dibujo/>, la cual se colocó dentro del etiquetado justo donde el niño puso el dibujo dentro de su texto.

Puede parecer que no sea muy importante señalar los dibujos de los niños, pero pensar eso es caer en un error. El hecho de que algunos niños dibujen muestra que tienen otro medio para expresarse, incluso en el trabajo de otros compañeros etiquetadores se dio el caso donde el niño hacia un dibujo para expresar una palabra o una idea, es decir que en lugar de escribir lo que quería decir lo dibujaba. Lamentablemente en el objeto de estudio de esta tesis no se encontró nada similar, sin embargo en este material la mayoría de los dibujos están relacionados con la lectura, no son arbitrarios; aunque los niños no sustituyen las palabras por los dibujos, hay algunos casos donde dibujan exactamente lo que describen en su texto, incluso en unos casos se dibujan a sí mismos al lado de su firma como se puede ver en la ilustración N.

4.4.7. Comentarios

Como explicamos anteriormente, al final de los textos se abrieron algunos espacios para escribir los comentarios, que más que eso son descripciones de fenómenos que no se pueden transliterar, como las letras encimadas, las cursivas, los adornos, dibujos, los diferentes colores de tinta, etc.

Todos estos comentarios fueron puestos desde la transliteración y nos pareció que era necesario conservarlos en el etiquetado, ya que nos ayudan a entender varias cosas que suceden pero que ya no podemos ver. De esta forma creamos las etiquetas **<comentarios></comentarios>**, las cuales marcan el principio y final de todo el espacio para los comentarios y dentro de este espacio se implementaron las etiquetas **<com></com>** entre las cuales encerramos cada una de las descripciones. Así, los espacios quedan de esta forma:

<comentarios>

<com>3. Bajo la "a" hay una "e".**</com>**
<com>4. En "mama" el acento está sobre la segunda "m".**</com>**
<com>11. Bajo la "E" hay una "e".**</com>**
<com>15. “Manuel” está escrito con letra cursiva**</com>**
</comentarios>

4.5. Validación del etiquetado

Con todo lo anterior podemos darnos cuenta que se utilizaron diferentes etiquetas en varias partes del texto, cada una de ellas con su función y reglas correspondientes, pero al haberse realizado la mayoría del etiquetado de forma manual es indudable que pueden haber algunos errores.

Para corregir cualquier error se recurrió a dos tipos de medida: La primera consiste en servirnos de las herramientas de validación del XML, de las cuales ya hemos hablado en el capítulo correspondiente a este metalenguaje. Con ellas se identificaron los errores consecuentes de la mala redacción, la omisión, la inserción o la mala ubicación de alguna etiqueta. Así, una vez que el XML identificó los errores se corrigieron manualmente.

Las herramientas del XML nos ayudaron a encontrar estos errores gracias a que identifican de manera automática si el etiquetado cumple con los dos tipos de reglas del XML. Como recodaremos, las primeras reglas consisten en que las etiquetas deben estar escritas y ubicadas en la forma como nosotros mismos especificamos en el esquema; y las segundas reglas son las generales, las que debe seguir cualquier etiquetado de XML sea lo que sea que se esté etiquetando. Estas últimas reglas residen en que toda etiqueta de apertura debe tener una etiqueta de cierre, las etiquetas que van solas usan la diagonal de diferente manera en comparación con las etiquetas que van en pares, y siempre hay que señalar la versión del metalenguaje y la dirección donde está ubicado el esquema.

Una vez que corregimos las etiquetas con ayuda del XML, nos percatamos que había errores del etiquetador que no identificaba la validación automática, estos errores fueron:

- Los etiquetadores olvidaron alguno de los nuevos convenios, por ejemplo que en “esteban” decidimos considerar una sustitución de mayúscula por minúscula y ya no una omisión de mayúscula.

- Olvidaron algún atributo o elemento que completaba la información, por ejemplo, a pesar de que añadieron las etiquetas que marcaban los fenómenos de una palabra, olvidaron poner el atributo “n” cuando era necesario. Asimismo, cuando un punto servía como herramienta de segmentación, algunas veces los etiquetadores olvidaron poner en el atributo “punto” la *s* que diferencia los puntos de segmentación de los ortográficos.

De esta forma, la segunda medida que tomamos fue intercambiar el material entre los etiquetadores para que cada uno revisara y corrigiera lo que el otro etiquetador había olvidado.

Así, después de haber desarrollado todo este capítulo podemos asegurar que, de acuerdo con los lineamientos seguidos, nuestro corpus es fiel a los textos de origen gracias a la digitalización, transliteración y normalización; completo y detallado, ya que explora distintos niveles de la lingüística que van desde el ortográfico, pasando por lo léxico y concluyendo en lo morfológico; y finalmente, confiable gracias a las múltiples verificaciones que se hicieron tanto automática como manualmente.

5. VISUALIZACIÓN DEL CORPUS DE DISCURSO INFANTIL ESCRITO

5.1. La hoja de estilo

Como se ha visto en los últimos capítulos, la construcción del corpus del proyecto *Aprender mientras se enseña: una experiencia de acompañamiento en la enseñanza de la Lengua escrita* se ha llevado a cabo a partir de varios procesos que van desde la elaboración del esquema hasta el etiquetado. Esto, con la finalidad de obtener una herramienta confiable que nos ayude a realizar varios análisis a nivel léxico y ortográfico en los textos de los niños. Pero además de eso, uno de los objetivos del proyecto es que el corpus sirva como material para diversas investigaciones en distintas áreas, por tanto, resolvimos crear un formato electrónico de fácil acceso que sea comprensible para cualquier persona que esté o no relacionada con el metalenguaje XML.

Para este corpus el Mtro. Carlos Francisco Méndez configuró el XSL, el cual, como recordaremos, es una herramienta del XML y consiste en diseñar un esquema con todas las indicaciones necesarias para que en la hoja de estilo se exponga la información que queremos con la presentación que consideremos más adecuada.

De esta forma obtuvimos la siguiente hoja de estilo donde se exhibe todo lo que el usuario pudiera requerir del corpus:

NIÑO	
Nombre:	Irma [redacted]
Sexo:	Femenino
Edad:	7/10
Fecha de nacimiento:	1998[redacted]
Grado:	2
Grupo:	A
MAESTRO	
Nombre:	Emma [redacted]
DIRECTOR	
Nombre:	Javier [redacted]
ESCUELA	
Nombre:	[redacted]
Clave:	01
Zona:	[redacted]
Sector:	[redacted]
ACOMPAÑANTE	
Nombre:	Francisco [redacted]
ARCHIVOS	
Trasliteración:	N[redacted]0.txt
Escaneo (imagen):	N[redacted]0.jpg

1. MI escuela
2. ¿Quin soy<?> IRMA [redacted]
3. y estoy en 2 grado y tengo 3 Amigas y 7 amigos me gusta comer
4. tengo 7 años ymis amigas se llaman leslie, Belen y Lupita
5. y amigos jose , juan , carlos ,Ruben ,Alexis, Luis y marcos mi rancho
6. sellama el [redacted] ymi pais se <llaman> llama Mexico mi casa es

7.de ladrillo y <de> me gusta comer lo quemi mam[á] me ase <y> tengo un p
errito

8. quese llama Pirolais
9. ¿Como es nuestra escuela? es bonita sellama [redacted] es de
10. Color verde tiene Lus electica , telebision , betila<s>ion tiene arboles
11. la puerta es de fiero tiene libros al cancha tiene pisaron
12. tienen 6 salones y <ende> el piso esve sem ento tiene 2 cuartitos
13. uno de apo<y>o y la tiendita y la direcsion tiene <6>años
14. <como llem ono ala>
15. ¿como llegamos ala escuela? miescuela esta un poco le gos y
16. mi mamá me trai y cuando me boy amicasa me boy sola
17. ¿Que semos en la escuela? mi mamá ase el ase o cuando
18. me toca leemos escribimos en cuader<n>os y libros y salimos
19. a e<d>ucasion <f>isica en <e>ducasion fisi<c>a jugamos y nos dibertimos
20. y tambien en el recreo adios

Comentarios:
No hay espacioes entre los párrafos, sino líneas.
1. Está centrado el renglón.
2. Bajo el "?" hay un "2".
6. Bajo la "R" hay una "r".
7. Las segundas "a" de las palabras "mamá" están acentuadas con triángulos;
Bajo la "y" hay una "i" o "l".
10. Bajo la "s" hay una "c".
13. Bajo la "y" hay una "ll"; Bajo el "6" hay una "a".
18. Bajo la "n" hay una "m".

TABLA RESUMEN	
Total de palabras (<g>):	204
Mayúscula inicial:	4
Mayúscula intermedia:	3
Hipersegmentación:	2
Hiposegmentación:	12
Omisión e:	1
Omisión acento:	22
Omisión h:	2
Omisión r:	3
Omisión n:	1
Omisión ha:	1
Sustitución ixI:	2
Sustitución rxR:	1
Sustitución mxM:	1
Sustitución axA:	3
Sustitución nxN:	1
Sustitución gxG:	1
Sustitución exE:	1
Sustitución lxL:	2
Sustitución cxC:	2
Sustitución Yxy:	1
Sustitución Txt:	1
Sustitución Lxl:	1
Sustitución Jxj:	2
Sustitución Cxc:	2
Sustitución Mxm:	1
Sustitución cxs:	8
Sustitución zxs:	2
Sustitución vxb:	5
Sustitución dxv:	1
Sustitución jxg:	1
Sustitución exi:	1
Permutación l:	1

- Como se expuso anteriormente parte del etiquetado está formado por un encabezado con toda la información que tenemos sobre el origen de los textos. Esto se hizo debido a que la información socioeconómica de los niños puede ser relevante para algunas investigaciones, ya que que el aprendizaje se da y el conocimiento se adquiere de distintas formas en cada nivel social. Es así como al principio de la hoja de estilo aparece esta información como se muestra en la ilustración O.

- Posteriormente vemos el texto del niño en forma transliterada, sólo que a diferencia del formato txt, el formato que nos ofrece el XSL es didáctico, ya que al colocar el mouse sobre alguna palabra donde se nota alguna anomalía, aparece una ventana donde se explica qué es lo que sucede con la palabra en cuestión, por ejemplo en la imagen P al indicar con el mouse la palabra “betila<s>ion” se abre una ventana que nos describe las sustituciones, omisiones y corrección que aparecieron.

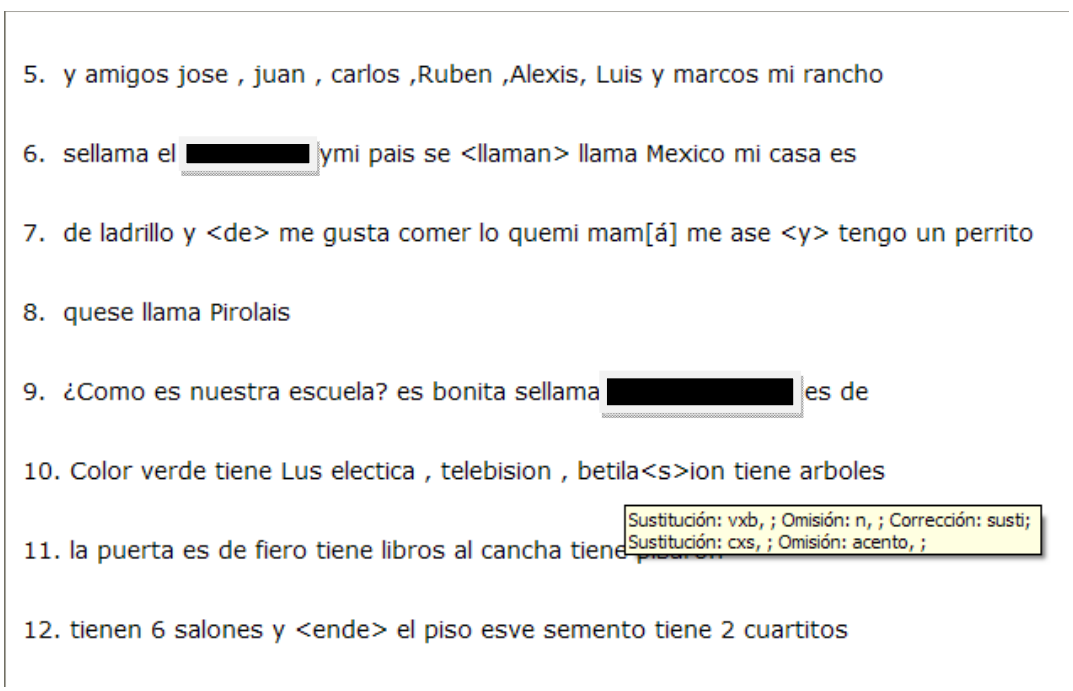


Ilustración P. Ventana de descripciones.

- En seguida del texto encontramos los comentarios, que más que eso son descripciones de fenómenos que no se pudieron transcribir como letras encimadas, letras cursivas, adornos, subrayados, etc.

- Por último podemos ver que la hoja de estilo nos proporciona una tabla en donde de manera automática se muestra la suma de absolutamente todos los fenómenos que se encontraron, analizaron y etiquetaron.

5.2. Análisis cuantitativo

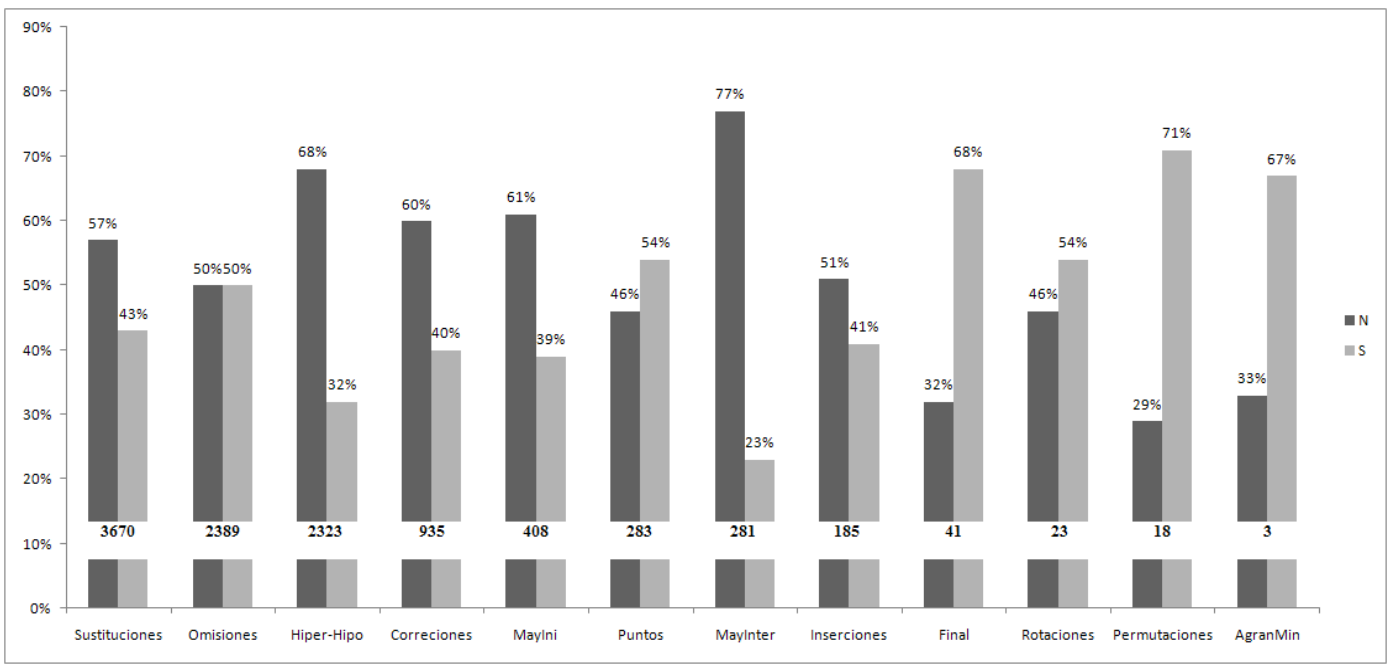
Como se acaba de ver en el punto anterior, la hoja de estilo está configurada para hacer conteos automáticos de las etiquetas que se utilizaron. De esta forma todo el trabajo que se describió en el capítulo cuatro se hace evidente en estas tablas de resumen, pero cada una de éstas sólo muestra la suma de fenómenos del texto al que pertenecen, y aunque esto es información muy valiosa, no nos aclara del todo muchas dudas que se fueron formulando en capítulos anteriores. Por tanto, decidí extraer manualmente todos los resultados que arroja cada tabla de resumen para hacer gráficas comparativas que nos ayuden a ver más claramente qué fenómenos fueron los más recurrentes tanto en cada escuela como en general. De esta manera, en este punto haré un análisis cuantitativo de cada uno de los fenómenos con base en las gráficas comparativas que hice.

Es importante aclarar que las comparaciones que se expondrán en este apartado se han hecho con base en el objeto de estudio de la tesis, asimismo, los análisis que expondré no fueron hechos a solicitud del proyecto. Éstos los hice con el fin de de utilizar mis conocimientos obtenidos en la universidad y entender de la mejor manera las posibles causas de los fenómenos antes explicados, por supuesto debo aclarar que mis conocimientos no aclararán del todo algunas dudas, para ello es necesario especializarse en ello. Asimismo, es importante que quede claro que aquí sólo se expondrán los resultados obtenidos del objeto de estudio y éstos no deben generalizarse para el resto del material del proyecto.

Tomando en cuenta que los niños que más escriben son propensos a tener más errores, en las gráficas comparativas hice *porcentajes de error*, es decir, con base en el número de palabras que escribió cada niño obtuve los porcentajes que ocupan los fenómenos en cada uno de los textos, por ejemplo, si el niño A tuvo 6 omisiones en un rango de 24 palabras y el niño B tuvo 3 omisiones en un rango de 10 palabras, el niño A

tiene menos omisiones que el niño B, ya que las omisiones en el niño A ocupan el 25% del texto y en el niño B ocupan el 30%.

Una vez que extraje los resultados de cada una de las tablas de resumen, clasifiqué doce grupos principales: Hiper e hiposegmentaciones, mayúsculas iniciales, mayúsculas intermedias, agrandamiento de minúsculas, sustituciones, rotaciones, omisiones, permutaciones, inserciones, correcciones, puntuación y términos de líneas. Posteriormente, en cada uno de estos grupos hice una suma del total de las apariciones, dando como resultado la gráfica 1:



Gráfica 1 Fenómenos principales

En esta gráfica se presentan los doce grupos en orden de acuerdo con el número de apariciones, el cual se ubica en medio de su barra correspondiente. Asimismo, señalé el porcentaje de error que tuvo cada conjunto de escuelas, que como recordaremos el conjunto N se refiera a las escuelas cuyos maestros no participaron en las capacitaciones, y el conjunto S se refiere a las escuelas cuyos maestros sí participaron en las capacitaciones.

De esta forma podemos ver que los tres fenómenos más frecuentes en los textos de todos los niños son las sustituciones, las omisiones y la mala segmentación de las palabras, y dentro de estos tres fenómenos el conjunto N ocupa el mayor índice de

porcentaje, es decir que los niños que pertenecen a un grupo donde el maestro no fue capacitado son más propensos a cometer estos tres tipos de errores.

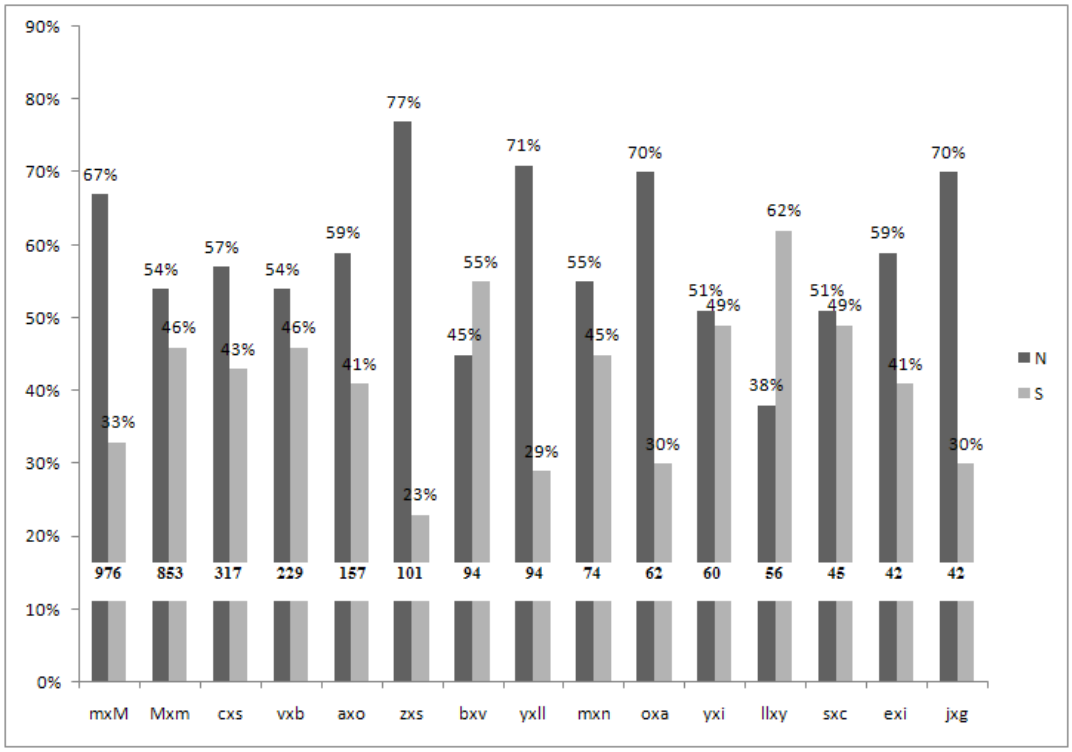
Hay que notar además que el conjunto N tiende a corregirse más que el conjunto S, pero después de ver las primeras tres barras podemos suponer que el hecho de que el conjunto N se corrija más, deriva de que tiene más errores, lo cual nos hace dudar si su porcentaje de corrección es bueno o en realidad es malo, ya que significa que en tal caso hubo muchos más errores. En cambio el porcentaje de correcciones para el conjunto S adquiere un carácter positivo, ya que además de que tiene menos porcentajes de error tiene un buen porcentaje de correcciones.

Con respecto a la puntuación podemos notar que ambos conjuntos hacen buen uso de ella, y aunque el conjunto S aventaja al conjunto N, realmente no hay mucha diferencia. Pero sí existe diferencia con respecto a la forma en cómo terminan los renglones, ya que a pesar de que en todo nuestro corpus hay pocos fenómenos a final de renglón, casi todos estos pertenecen al conjunto S. Sin embargo, esto no se puede considerar una desventaja para el conjunto N, simplemente los niños del conjunto S tienen un conocimiento un poco más concreto sobre las reglas básicas de la redacción.

A pesar de que esta gráfica nos ha mostrado información muy valiosa aún quedan algunas incógnitas que sólo pueden resolverse con información más detallada. Por tanto, realicé un análisis de cada uno de los doce grupos.

5.2.1. Sustituciones de mayúsculas y letras

Siguiendo el orden de mayor frecuencia comenzaremos con las sustituciones, las cuales como ya hemos dicho pudieron ser consecuentes por la similitud fonológica que el español de México le da a algunas grafías como *b* y *v*, por errores de concordancia como *casa blanco* o por la similitud gráfica como *h* y *n*. Pero estas son meras conjeturas que de ser ciertas encontraremos muchos ejemplos similares y, lo más importante, veremos si hay o no un patrón.



Gráfica 2 Sustituciones

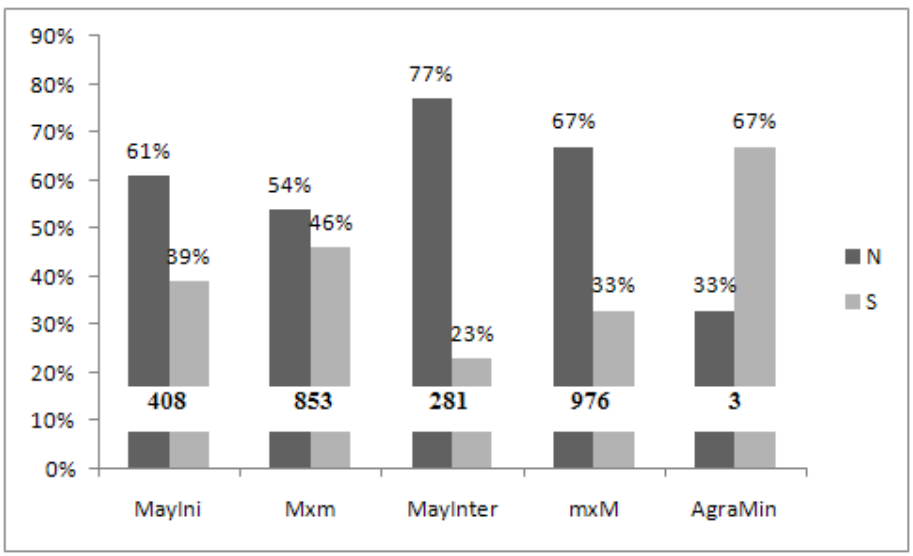
En la gráfica 2 se muestran sólo las sustituciones con mayor frecuencia prescindiendo de las de menor, ya que estas últimas, al ser casos muy singulares, no representan un patrón general.

En la parte superior de cada barra podemos ver el porcentaje de sustituciones que hizo cada conjunto de escuelas, un poco abajo se encuentra el número total de apariciones de cada sustitución y en la parte inferior vemos cuáles fueron las sustituciones más usuales.

Las primeras dos barras nos indican las sustituciones que hubo entre mayúsculas y minúsculas, y como puede verse éstas obtuvieron el índice más alto de aparición. Si recordamos el uso de las etiquetas, podemos entender que el error más recurrente de los niños es sustituir las minúsculas por las mayúsculas, y como puede verse el conjunto N supera por mucho al conjunto S, lo cual podría indicarnos que a los niños del conjunto N se les dificulta aprender la grafía de las minúsculas o no saben con exactitud cómo se deben usar las mayúsculas.

Posteriormente tenemos que los niños suelen sustituir las mayúsculas por minúsculas, pero a diferencia de los resultados anteriores nos damos cuenta que la diferencia entre los conjuntos no es mucha, lo cual podría indicarnos que ambos

conjuntos entienden que los nombres propios, las palabras a principio de párrafo o antecedidas por un punto usan mayúsculas iniciales. Pero con respecto al conjunto N, esta última deducción se contradice con la del párrafo anterior, así que para complementar esta información incluimos el análisis del uso de las mayúsculas iniciales y las mayúsculas intermedias con la siguiente gráfica:



Gráfica 3 Uso y sustitución de mayúsculas

Como puede verse en esta gráfica, no se ordenaron las barras por número de aparición, sino que primero se puso la barra que representa uno de los cumplimientos de las normas de escritura, el uso de mayúscula inicial. Seguida de ésta encontramos su contraparte que es la sustitución de mayúsculas por minúsculas. Posteriormente vemos una transgresión que consiste en colocar mayúsculas en medio de una palabra, después está su contraparte que es la sustitución de minúsculas por mayúsculas y finalmente vemos el fenómeno de agrandamiento de minúsculas. Este último lo puse aquí porque no sabemos si una minúscula agrandada puede considerarse como una mayúscula.

Retomando las deducciones de los párrafos precedentes a esta última gráfica, podemos notar que aunque el conjunto N tiene un mayor porcentaje de uso de mayúsculas iniciales, también tiene una mayor porcentaje de uso de los últimos dos fenómenos. ¿Qué nos dice esto? que en el conjunto N la mayoría de los niños hace uso excesivo de las mayúsculas, es decir que no sólo usan las mayúsculas como se debe, sino que además reparten mayúsculas en sus textos, lo cual nos indica que a pesar de

que tienen mayor porcentaje de uso de mayúsculas iniciales realmente no saben hacer uso apropiado de ésta.

En contraparte podemos ver que a pesar de que el conjunto S tuvo un porcentaje menor de uso de mayúsculas iniciales, también tuvo un índice menor en todos los demás fenómenos, lo cual nos indica que los niños de este conjunto hacen un uso más proporcionado de las mayúsculas y las minúsculas.

Con respecto a las minúsculas agrandadas analizamos los tres ejemplos encontrados y todos coincidieron con que la minúscula agrandada está a principio de palabra, es decir que podría tener función de mayúscula, pero de ser así entonces en dos de los ejemplos se señalaría una sustitución de minúscula por mayúscula ya que la sintaxis de la oración no permite que haya mayúscula como se podrá ver:

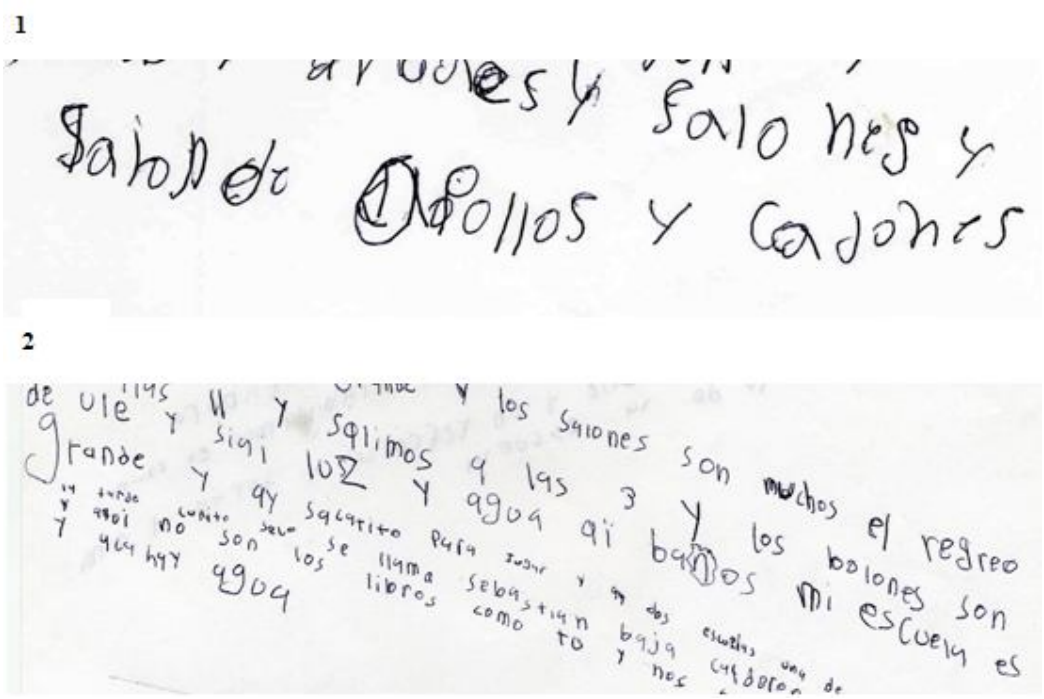


Ilustración Q. Agrandamiento de minúsculas I.

En cambio con el tercer ejemplo encontrado, si atribuimos a la minúscula agrandada como mayúscula la consideraríamos como mayúscula inicial de la palabra debido a que ésta está a principio de renglón e inicia con una idea diferente al párrafo anterior.

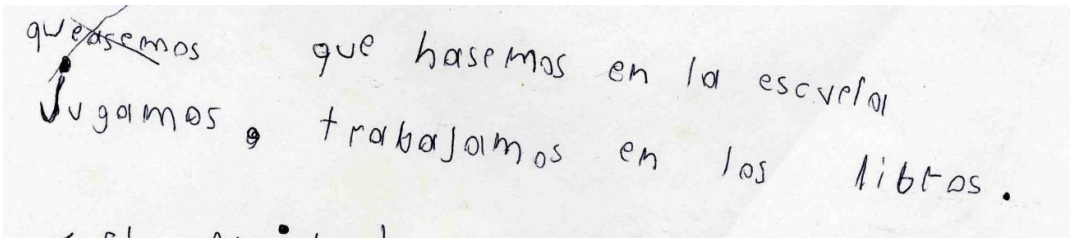


Ilustración R. Agrandamiento de minúsculas II.

Enseguida de las sustituciones entre mayúsculas y minúsculas encontramos en la gráfica 2 dos patrones que bien pueden deberse a la similitud fonológica que el español de México le da a las grafías, hablamos de las sustituciones de la *c* por la *s*, y de la *v* por la *b*. Nuevamente el conjunto N tiene un mayor índice de error pero no por mucho. ¿Qué podemos deducir de esto? en el caso de la *c* y la *s* podemos darnos cuenta que el sonido /s/ está más relacionado con la *s* ya que las sustituciones de la *s* por la *c* tienen un número de aparición mucho menor. Con respecto a la *v* y la *b* podemos suponer que, además de que el sonido /b/ se relaciona más con la *b*, los niños tienden a aprenderse más la grafía *b* posiblemente porque está al inicio del abecedario.

Con respecto a las sustituciones de las *a* por las *o* no podemos suponer que se deba a un problema fonológico, más bien se podrían deber a otras dos posibles causas:

- La similitud gráfica. Durante el aprendizaje del discurso escrito los niños encuentran la forma que les parece más cómoda para escribir las letras y con respecto a la *a* una de esas formas puede ser escribir primero una *o* y luego poner una raya para que se convierta en *a*, de esta forma el olvido de una aparente simple raya nos da los resultados que vemos.
- La mala concordancia morfológica entre artículos, sustantivos, verbos, etc. En la mayoría de las palabras la forma simple termina con *o*, principalmente adjetivos, por lo que muchos errores de concordancia pueden deberse al olvido o a la ignorancia de que en ocasiones la *o* debe ser sustituida por una *a* para que la relación entre los elementos de una oración sea correcta.

Con respecto a las demás sustituciones que se muestran en la gráfica, las causas que las provocan pueden ser iguales o parecidas a las que ya hemos supuesto, pero las sustituciones de *e* por *i* no encajan en éstas, ¿Qué es lo que probablemente sucede entonces? Recolectemos algunos ejemplos del objeto de estudio:

1. Qué **A ciMoSen** laescuRla *(hacemos en)*
2. y los **binimos** a Pie ytabien la poto *(venimos)*
3. mevengo go apie **aciorentor** *(a qué hora entro)*
4. ladriyo con fiero ay escobas **trapiadores** *(trapeadores)*
5. ami **me ta ay metrai** mi papa *(me trae)*

Como podemos darnos cuenta ocurren ciertos fenómenos fonológicos que han acompañado a la lengua por siglos, nos referimos a la *disimilación, asimilación* y *sinéresis.* La primera consiste en que "un sonido en posición privilegiada" desemeja "a otro en no tan privilegiada situación" (Bolaño: 1968, p. 143), un ejemplo de esto es el primer enunciado.

En el triángulo vocálico podemos notar que la *e* está muy próxima a la *a* en el punto de articulación, por tanto, la diptongación de estos fonemas puede ser "difícil" [38] de pronunciar, es así como inconscientemente las personas encuentran formas para hablar más fácilmente[39], y una de estas formas es la disimilación. Así, la *e* se sustituye por la *i*, con la cual comparte la característica de la palatalización, pero ésta está más lejos de la *a*.

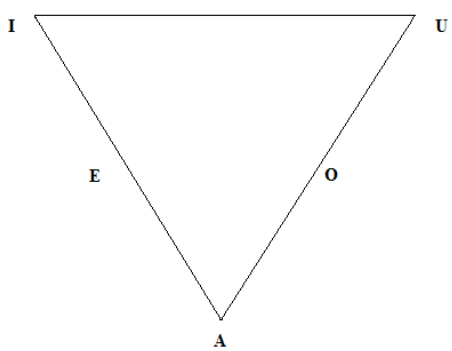


Ilustración S. Triángulo vocálico.

Por el otro lado encontramos el fenómeno de asimilación del cual también nos servimos para facilitar nuestra comunicación oral y consiste en que "unos sonidos convierten a otros de la misma palabra en semejantes" (Bolaño: 1968, p. 143). Un ejemplo de esto sería el enunciado número dos donde la *i* influyó en la *e* asimilándola totalmente.

Finalmente tenemos tres casos con sinéresis, pero para entender de qué se trata este fenómeno primero vamos a explicar qué es un hiato y un diptongo:

Un hiato se presenta cuando encontramos: a) dos vocales juntas siendo una de ellas alta (i, u) y la otra media o baja (e, a, o); b) dos vocales juntas siendo ambas medias (e,o), o una de ellas media y la otra baja (a). Sea cual sea el caso cada una de las

[38] La diptongación entre la *a* y la *e* es difícil en comparación con la diptongación de la *a* con la *i*.
[39] Como dice Alarcos "Esto sucede en virtud de la *tendencia económica* del sistema, fuerza de estructura paralela a la ley del mínimo esfuerzo y de la inercia en el habla". 1965. Pág. 123.

vocales corresponde a una sílaba diferente, por ejemplo: ha-c**í-a**, con-ti-n**ú-o**, **a**-h**í**, cé-r**e-o**, a-s**e-o**, s**o-e**z, etc.

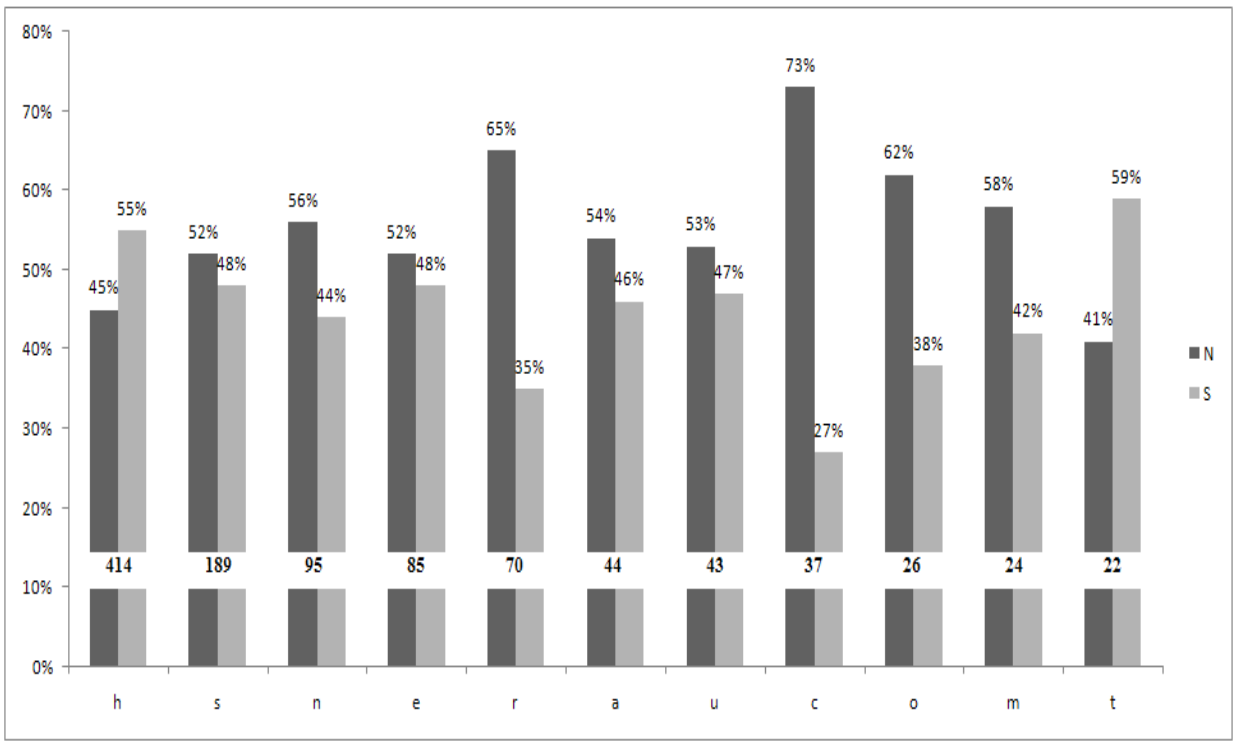
Un diptongo se presenta cuando encontramos juntas dos vocales átonas, formando entonces en conjunto una sola sílaba, por ejemplo: ha-c**ia**, con-ti-n**uo**, ra-d**io**, etc.

Ahora bien, encontramos una sinéres cuando las vocales que forman sílabas diferentes se consideran en una sola sílaba, es decir, hay sinéresis cuando un hiato es considerado como diptongo, por ejemplo, en vez de segmentar cé-r**e-o** se segmenta cé-r**eo**. De la misma forma, en los ejemplos 3, 4 y 5 en vez de tener dos sílabas[40] ***que-ho****ra*, *tra-p****e-a****-do-res*, *tr****a-e*** tenemos una sola *c****io****r*, *tra-p****ia****-do-res*, *tr****ai***. Y no sólo eso, sino que además en los tres ejemplos hubo una disimilación de la *e* con lo cual ésta perdió un poco de su tonicidad y la sinéresis se dio de forma más fluida.

5.2.2. Omisiones

Como ya hemos dicho anteriormente hay letras en el discurso escrito que se omiten en la oralidad por varias razones, ya sea porque la letra carece de un sonido, porque en la oralidad estamos acostumbrados a omitir los últimos fonemas de algunas palabras, porque juntamos los sonidos cuando una palabra termina con el mismo fonema con el que comienza la palabra que le sigue, etc. Hay muchas razones por las que se pueden omitir las letras, por tanto, con la siguiente gráfica veremos que letras fueron las más omitidas y a partir de ahí analizaremos las causas.

[40] Que además en el ejemplo 3 cada una de ellas corresponde a palabras diferentes.



Gráfica 4 Omisiones principales

Al igual que en la gráfica 2 y en todas las que siguen, mostraremos únicamente los fenómenos donde hubo mayor cantidad de ocurrencias, ya que como hemos mencionado, sólo estos fenómenos pueden darnos patrones de error sustentables.

Ahora bien, como se puede ver en la gráfica anterior, la *h* es la letra que más se omite en los textos, lo cual no nos sorprende porque su carencia de sonido causa algunos problemas incluso a los adultos, por tanto, es de esperarse que los niños, los cuales están apenas aprendiendo a escribir, tengan dificultades con ella.

En seguida de la *h* encontramos las omisiones de la *s* y la *n*, las cuales, al revisar los ejemplos del corpus, nos dimos cuenta que en general son causadas por una mala concordancia de número, por ejemplo:

- entramos**Alaocho** tempranito *(a las ocho)*[41]
- manuel y tengo **muchos primo y prima** *(muchos primos y primas)*
- me**gusta**lasmatematicas *(gustan)*

[41] Cabe la posibilidad de que este ejemplo no nos muestre una mala concordancia de número, porque esa omisión de *s* podría deberse a que el niño está relacionando el artículo *la* con el *8*, el cual es singular. De ser este el caso el niño no hace una mala concordancia de número, simplemente ignora que el verdadero sustantivo son las *horas* y que el artículo debe relacionarse con éstas.

- y llotengo anigos que se **llama**carlos y *(llaman)*

Por supuesto podemos encontrar excepciones donde los niños simplemente olvidaron poner las grafías como en los ejemplos árbole *(árboles)* y co *(con)*, pero en los ejemplos tego *(tengo)*, tramos *(entramos)*, tiedita *(tiendita)*, conocimito *(conocimiento)* y guta *(gusta)* no podemos considerar un simple olvido. Retomando la ley del mínimo esfuerzo, observemos que en todos los ejemplos las *n* y la *s* están a final de sílaba, posición donde es muy común que estos fonemas se asimilen al punto de articulación del fonema que les sigue, y en estos casos la asimilación se pudo dar a tal grado que la *n* y la *s* fueron poco perceptibles al oído y en consecuencia omitidas en la escritura.

Retomando el ejemplo de "conocimito" podemos notar que nuevamente la diptongación provoca alguna anomalía, en este caso fue la omisión de *e*. Como vimos en las sustituciones de *e* por *i*, la diptongación influye de tal forma que la *e* es sustituida por la *i*, pero en este caso al ser una diptongación entre una *i* y una *e* no podemos esperar que haya dos *i*, por lo que entonces la *e* se elimina.

Continuando con la *r* encontramos que la mayoría de las omisiones se dan en casos donde son precedidas por otra consonante, como por ejemplo "TenPano" (*temprano*). Este tipo de combinaciones fónicas son las últimas que aprende el niño a pronunciar, y como la enseñanza de la escritura no espera a que el niño termine por dominar la oralidad, muchos errores que comete en la oralidad podría cometerlos en la escritura también.

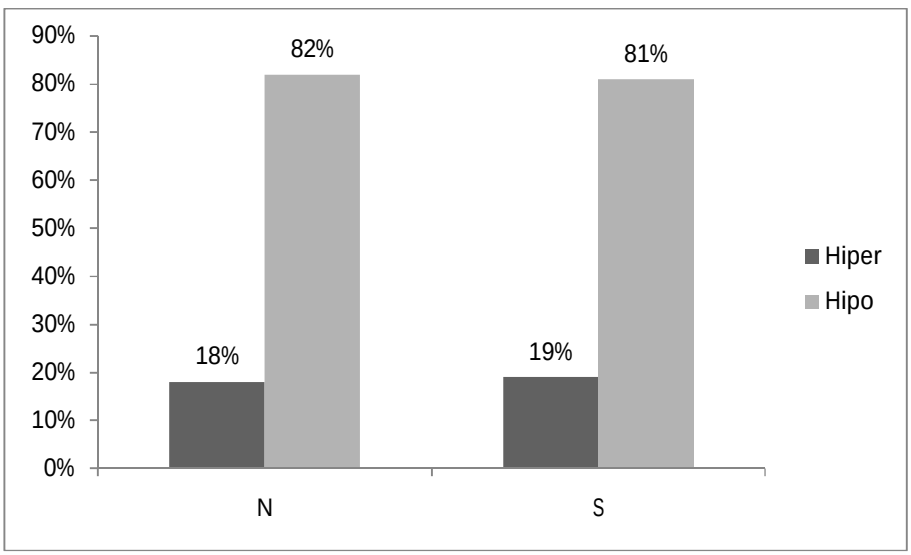
Con respecto a las *a*, la mayoría de las omisiones fueron arbitrarias, a excepción de algunos ejemplos donde tienen función de preposición, por ejemplo: "y nosotro nos gusta a ser" (y **a** nosotros nos gusta hacer), "JugamoALae chaAGararla chiquillas" (jugamos a la echa, **a** agarrar **a** las chiquillas). En realidad son muy pocos los ejemplos donde la *a* omitida coincidió con que era una preposición, así que no podemos especular nada sobre el manejo de esta preposición en los textos que conforman nuestro objeto de estudio, para ello se tendría que hacer otro tipo de análisis lingüístico que no realizaremos en esta tesis.

Con respecto a las omisiones de *c* y *o* notamos que no hay un patrón común, todos los ejemplos son diferentes entre sí, por tanto, concluimos que la omisión de estas dos grafías es arbitraria.

5.2.3. Hiposegmentaciones e Hipersegmentaciones

Como ya se ha mencionado antes, la correcta segmentación nos indica que los niños tienen claro la forma física de cada una de las palabras y el significado de éstas. Por tanto, podemos considerar que una mala segmentación se relaciona en varias ocasiones con una mala ortografía, ya que si el niño supiera la verdadera forma física de la palabra y su significado no encontraríamos ejemplos como "aser" cuyo significado sólo se puede desambiguar con ayuda del contexto, identificando así si el niño quiso decir "hacer" o "a ser".

Al ver la gráfica 1 podemos notar que el conjunto N tiene el mayor índice de error, pero dentro de ese resultado no sabemos qué porcentaje en los textos ocupa cada fenómeno, por tanto, desglosamos los resultados en la siguiente gráfica:



Gráfica 5 Hipo e hipersegmentaciones

Como puede verse la hiposegmentación ocupa parte muy importante en los textos, y tal parece que no sólo en estos, sino en muchos de los textos de escritores debutantes, ya que si recordamos los resultados del proyecto *La adquisición de la lengua escrita en diversos contextos lingüísticos y educativos* podemos ver que tanto en español como en portugués e italiano, dentro de los fenómenos de segmentación, la hiposegmentación es el más recurrente.

Una de las explicaciones que dieron en ese proyecto fue que el tipo de letra influye en la aparición de los fenómenos de segmentación, ya que los niños que escriben

en cursiva tienen menos errores de segmentación que los niños que no escriben en cursiva.

Aunque en el esquema del proyecto que enmarca esta tesis no hay etiquetas que detallen los fenómenos de segmentación, gracias a los estudios de Ferreiro y sus colegas (1996) sabemos que muchas de las hiposegmentaciones están constituidas por palabras donde una o varias de ellas se componen por una, dos o tres letras, es decir que muchas de las palabras que constituyen este fenómeno son artículos, pronombres átonos, preposiciones, etc.

Un fenómeno muy parecido ocurre en la oralidad: el sirrema. Éste es "la agrupación de dos o mas palabras que constituyen una unidad gramatical, unidad tonal, unidad de sentido, y que, además, forman la unidad sintáctica intermedia entre la palabra y la frase" (Quilis: 1999. p. 372).

No sabemos si las hiposegmentaciones de nuestro objeto de estudio se derivan también de alguna de las unidades que constituyen al sirrema, pero con el análisis que hicimos podemos notar que algunos constituyentes de la oración que forman a los sirremas también forman a las hiposegmentaciones:

1) Artículo y sustantivo: unPisaron, lascanchas, unavaca[42].
2) "El pronombre átono y el elemento que en la cadena hablada viene a continuación de él o al que se une" (Quilis: 1999. p. 373): amimegusta, mospone (nos pone), micasa.
3) Adjetivo y sustantivo: coloramarillo.
4) Sustantivo y complemento adnominal: puertasdefiero (fierro).
5) Los elementos constitutivos de las perífrasis o frases verbales: me gustacomer, estaecha.
6) Frases adverbiales: nadamas, seportanmuivien.
7) La conjunción y lo que introduce: amigosyamigasyllo, sandiaytunas, CeciliayLeopoldo.
8) La preposición con lo que le sigue: apie, porelcuadro, ensegundo.

Por supuesto, aunque existen algunas coincidencias, en definitiva hay muchos ejemplos de hiposegmentación que no tienen nada que ver con el sirrema, pero como ya hemos dicho las etiquetas de segmentación que usamos para etiquetar este corpus no

[42] Los ejemplos por supuesto fueron tomados de nuestro objeto de estudio.

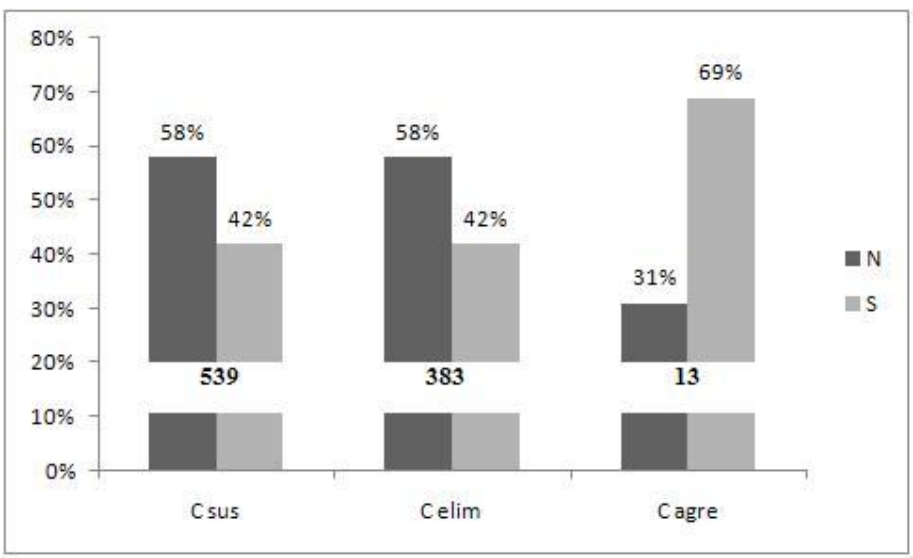
dan detalle de qué tipo de palabras hay alrededor de estos fenómenos, por lo que no podemos dar mayor explicación a la hiper e hiposegmentación.

5.2.4. Correcciones

Siguiendo con los fenómenos de la gráfica 1, encontramos lo que bien podría ser la contraparte de los puntos anteriores, las correcciones.

Aunque hasta ahora toda la información que hemos dado ha señalado que los fenómenos ocupan los primeros lugares en nuestro corpus, encontramos muy bien posicionadas a las correcciones, lo cual nos dice que a pesar de que no es usual, si es muy frecuente que los niños reflexionen sobre lo que han escrito.

Con la siguiente gráfica analizaremos con cuáles técnicas se valen los niños para autocorregirse.



Gráfica 6 Correcciones

Si recordamos las etiquetas de autocorrección podemos entender que la técnica más usada por los niños para corregir sus textos es la sustitución[43], la cual consiste en que los niños ponen la letra que consideran que está bien encima de la que creen que está mal, como se puede ver en la imagen T número 1, donde el niño encimó una *h* sobre una *n*.

[43] Csus: corrección por sustitución; Celim: corrección por eliminación; Cagre: corrección por inserción (agregar).

En segundo lugar vemos que los niños se sirven de eliminar por completo la palabra, frase o enunciado para posteriormente escribir la forma corregida o para cambiar de idea. En la ilustración T número 2 vemos que el niño tachó *mis papás*, posiblemente porque se dio cuenta de que ya había escrito algo semejante.

Por último, la gráfica 6 nos muestra que insertar las letras después de que se escribió la palabra es una técnica poco usada, ya que se encontraron sólo 13 textos donde por cada uno se localizó un ejemplo de esta técnica. Un ejemplo de inserción lo podemos ver en la imagen T número 3 donde es claro que el niño agregó una *r* después de que ya había escrito la palabra.

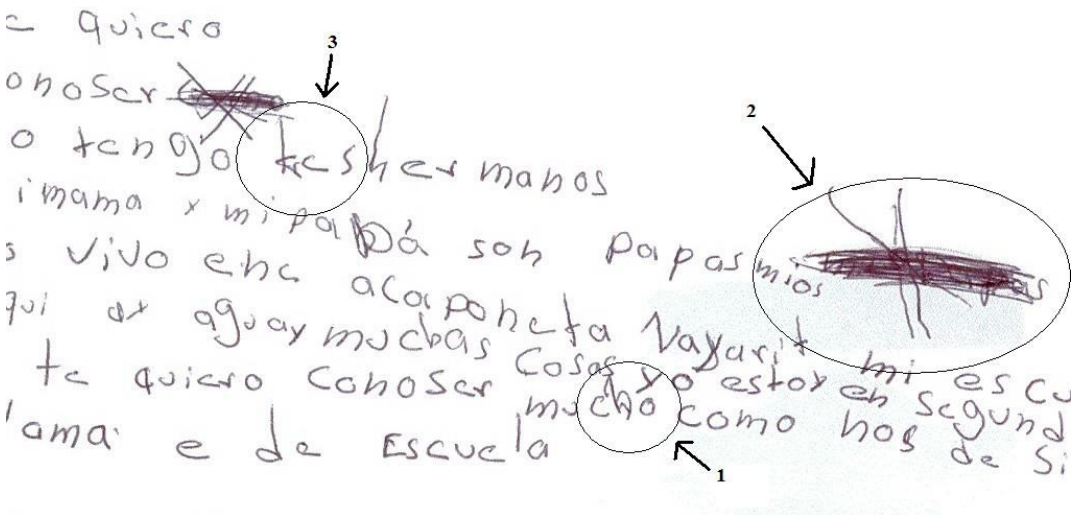
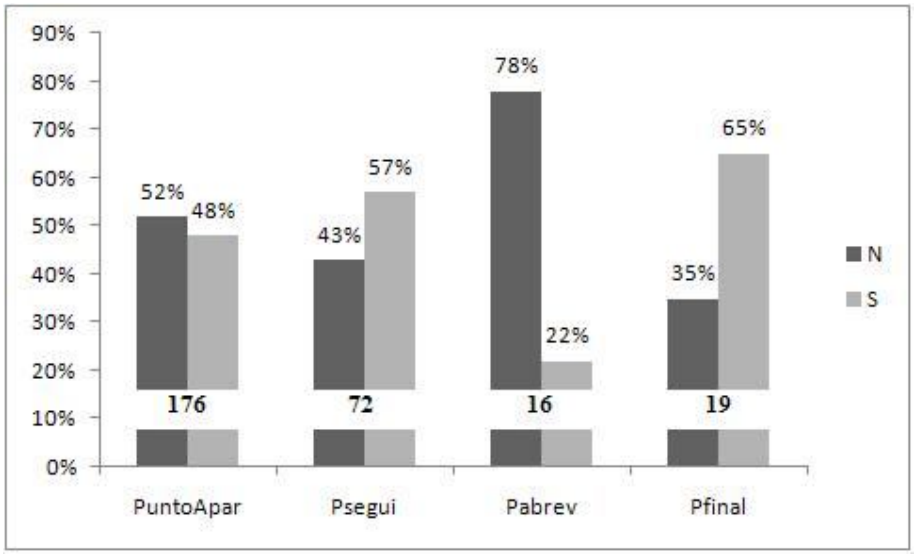


Ilustración T. Autocorrecciones: 1. Sustitución; 2. Eliminación; 3. Inserción.

5.2.5. Puntuación

Como ya se ha mencionado antes, la puntuación es muy importante para el discurso escrito debido a que ayuda a darle sentido al texto.

Al ver la gráfica 1 nos podemos percatar de que el uso de la puntuación sí es frecuente en el corpus, lo cual nos señala que los niños están aprendiendo de manera favorable la importancia que tiene la puntuación en los textos y la forma de uso de ésta. Por tanto, vamos a analizar cuáles signos usaron más y con qué frecuencia.



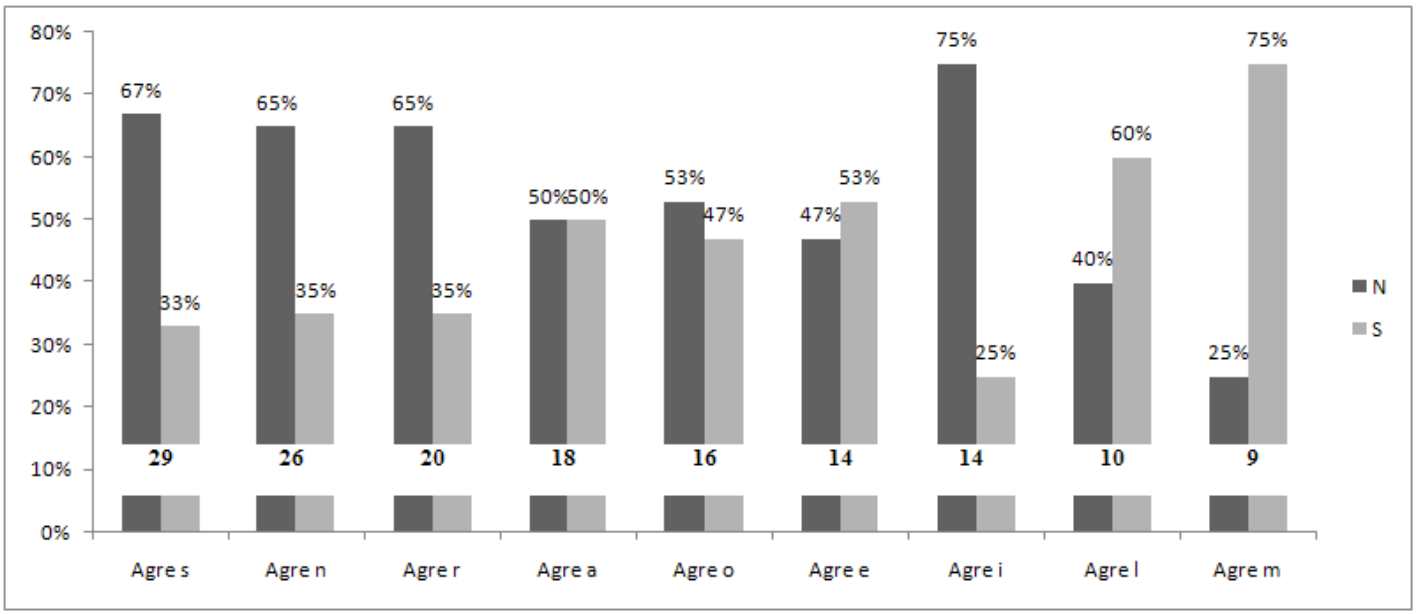
Gráfica 7 Puntuación

Como puede verse en la gráfica 7 la mayoría de los niños usan más los puntos que van a final de renglón[44], aunque curiosamente no usan mucho los puntos al final del texto. Asimismo, hay un porcentaje muy bueno del uso del punto y seguido.

5.2.6. Inserciones

Como ya habíamos mencionado en capítulos anteriores, dentro de todos los fenómenos de la oralidad que se presentan en la escritura encontramos también la inclusión de fonemas, como por ejemplo "cercas", "oyes", "haiga", etc. Pero como vimos, este fenómeno no siempre es provocado por la oralidad, sino que además hay otras razones las cuales recordaremos mientras analizamos la siguiente gráfica:

[44] PuntoApar: punto y aparte; Psegui: punto y seguido; Pabrev: punto de abreviatura; Pfinal: punto final.



Gráfica 8 Inserciones principales

Como puede verse la *s* y la *n* son las letras que más se insertan[45] en los textos, esto se debe a que en muchas ocasiones su uso puede hacer alguna diferencia de significado, es decir, en algunas palabras la inserción o la omisión de alguna letra no implica gran problema, como por ejemplo “escula” o “parrque”, pero en muchas palabras la inserción u omisión de una *n* o una *s* sí puede hacer una gran diferencia; por ejemplo, no es lo mismo “tiene” que “tienes” o “tienen”, y aunque en la mayoría de los niños es más común la omisión antes que la correcta concordancia, en algunos casos encontramos lo contrario: los niños agregan estas dos letras en palabras que no las requieren creyendo que así hacen una correcta concordancia. En el ejemplo “las bancas están forradas de *rositas*”, aquí el niño se dejó llevar por el uso de plurales, y la omisión del sujeto del complemento adnominal contribuyó a ello provocando que lo que en un principio era un adjetivo se volviera un sustantivo con un significado muy diferente, ya que en lugar de entender que las bancas están forradas de color rosa, entendemos que están forradas con rosas (flores).

Similar a esta situación encontramos muchos ejemplos donde el niño hace una buena concordancia en genero y número, sólo que no lo hace con la palabra correcta, por ejemplo en “hay una maestra que se llevan a los niños” el niño relacionó el verbo con “los niños” en vez con “maestra”.

[45] Agre = Inserción de…

Con respecto a las *r*, tal parece que los niños encasillan los dos fonemas de la *r* en dos grafías: la grafía *r* representar el sonido vibrante simple y la grafía *rr* representar al sonido vibrante múltiple. Como consecuencia de esto pudimos encontrar muchos ejemplos como “rregureo” (*recreo*), lo cual nos indica que, al encasillar los fonemas, los niños olvidan que a principio de palabra el fonema vibrante múltiple se representa con una sola *r*.

Posteriormente encontramos las vocales, y aunque tienen un porcentaje considerable de apariciones, no encontré un patrón de ocurrencias, por lo que concluyo que la inclusión de éstas es arbitraria, al igual que las inserciones de la *m*.

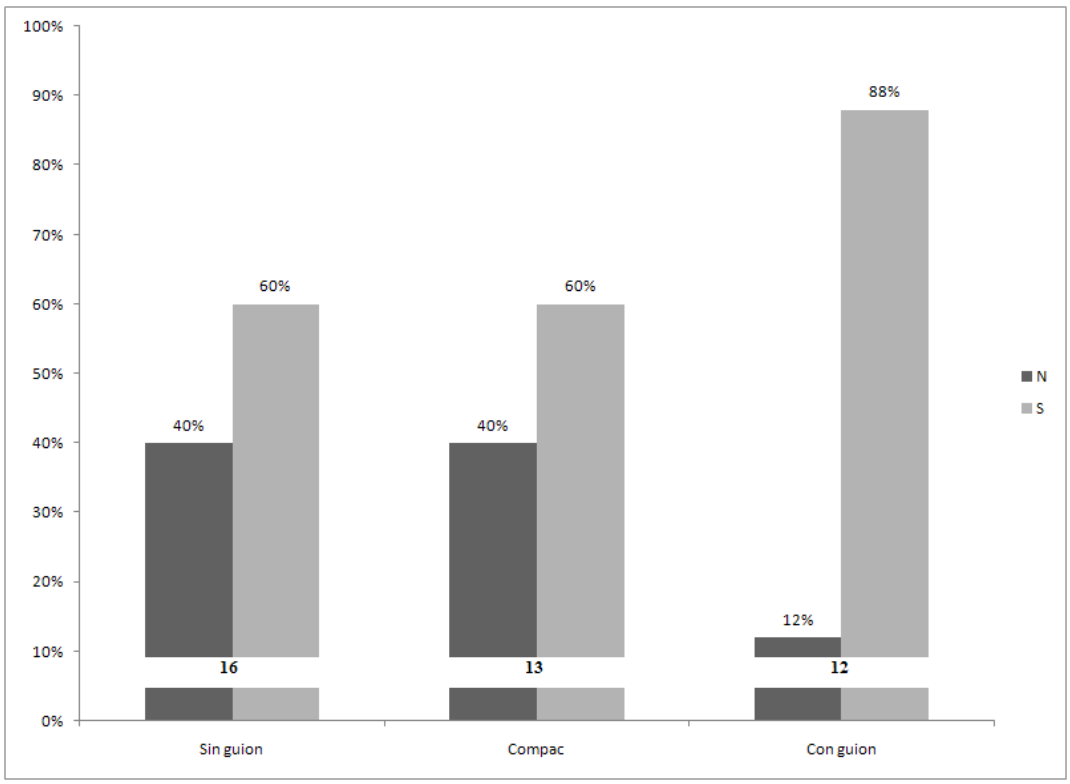
En relación con la inserción de las *l* podemos considerar que muchas apariciones fueron arbitrarias también, pero en algunos casos tal parece que los niños se dejan llevar por la existencia de la *ll*, ya que muchas veces encontramos esta grafía con valor de *l*, como en el caso de “escuella”.

5.2.7. Finales de renglón

Aunque la mayoría de los fenómenos que analizamos son a nivel gráfico, también nos importa mucho analizar las herramientas de las que se valen algunos niños para darle una estructura a su texto, como signos de puntuación, acentos, subrayados, títulos centrados, etc. Esto nos interesa debido a que, como se ha mencionado antes, consideramos que los niños que se sirven de más herramientas tienen un mayor conocimiento de las reglas básicas de la redacción.

Así, como parte de estas herramientas nos dimos a la tarea de analizar el uso del guión a final de renglón cuando se separa la palabra, ya que aunque ésta es la regla, los niños que la desconocen optan por otra solución.

En la siguiente gráfica veremos cuáles otras soluciones fueron encontradas y etiquetadas, y cuál fue el porcentaje de cada una:



Gráfica 9 Finales de renglón

A pesar de que en el etiquetado señalamos cuatro etiquetas, en todo el objeto de estudio sólo se encontraron tres.

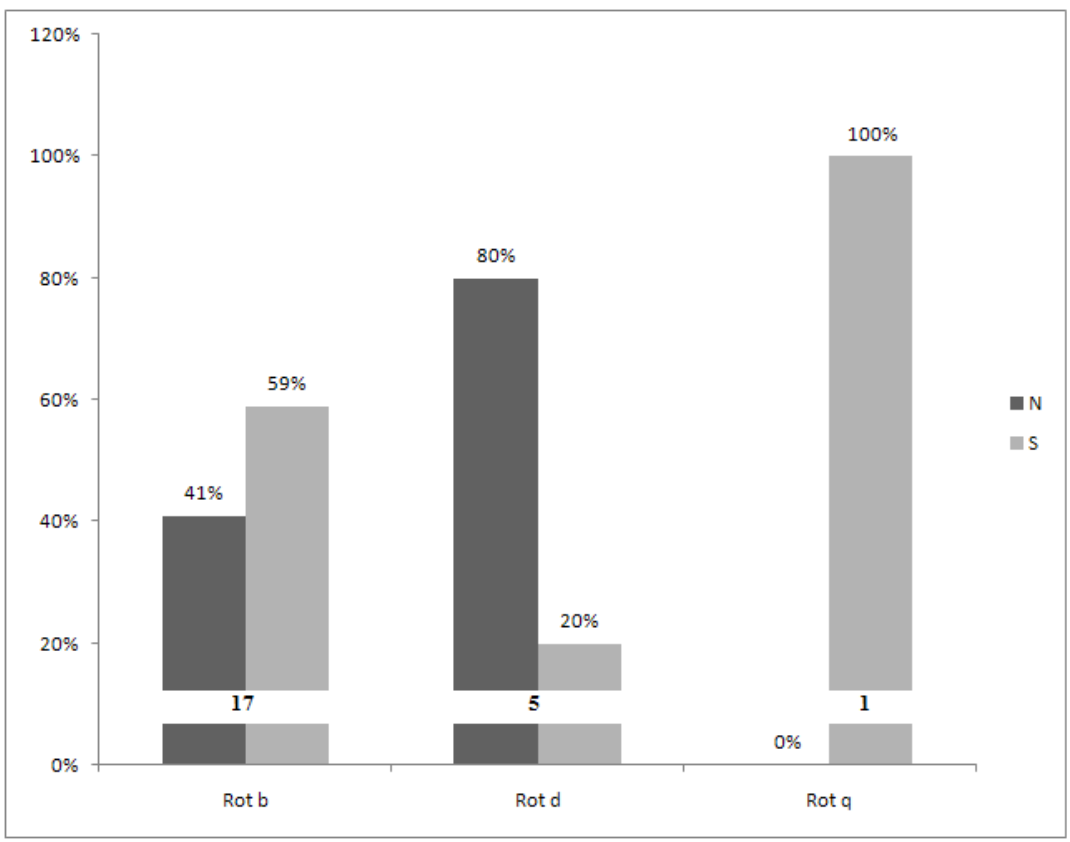
Si recordamos las etiquetas que se explicaron en el capítulo cuatro podemos entender con la gráfica que cuando la última palabra del renglón parece que no va a caber en el espacio que queda, lo que más hacen los niños es escribir una parte de la palabra y en el siguiente renglón la completan. En caso de que olviden o no conozcan esta técnica, la gráfica nos dice que entonces los niños optan por juntar o compactar las letras de esa palabra para que quepa. Con esto podemos ver que la mayoría desconoce la regla correcta, y que son muy pocos los que la ponen en práctica, ya que como se puede ver en la gráfica, esta regla convencional quedó en último lugar.

Asimismo, hay que notar que en todos los casos el conjunto S tiene mayor índice de ocurrencias, pero como ya mencionamos anteriormente esto no significa una desventaja para el conjunto N, ya que los niños de éste se valieron más de escribir las palabras mientras cupieran, y en caso de que el espacio no fuera suficiente, simplemente lo dejaban y continuaban escribiendo abajo.

5.2.8. Rotaciones

Muchos de los fenómenos antes mencionadas se deben a que el niño no ha aprendido del todo las reglas ortográficas, y aunque las rotaciones pueden ser consecuencia también de eso, podría ser que lo que realmente sucede es que el niño sabe qué sonido corresponde a tal o cual letra, sólo que a veces se confunde con algunas grafías por su similitud física, por ejemplo la *b* y la *d*, aunque sabe que a éstas les corresponde fonemas muy diferentes.

Afortunadamente, como podemos ver en la gráfica 1, este no es un fenómeno muy frecuente entre los niños, pero aún así desglosaremos los resultados para conocer cuáles fueron las grafías con las que se confundieron algunos niños.



Gráfica 10 Rotaciones

Como podemos ver, la *b* y la *d* encabezan la pequeña gráfica, en la cual lo que más llama la atención es que el conjunto S fue quien obtuvo mayor índice de rotaciones de *b*.

5.2.9. Permutaciones

Finalmente encontramos este último fenómeno, la permutación, que como se recordará consiste en que los niños usan las letras correctas pero en orden equivocado, por ejemplo “cosntruyo”.

Encontré pocos ejemplos de este fenómeno debido a dos razones: la primera es que los niños utilizaron, por supuesto en su mayoría, palabras muy sencillas donde las sílabas son simples y no compuestas por varias consonantes. Y la segunda razón es que las pocas palabras complejas que se encontraron tenían otros fenómenos.

Ahora bien, a pesar de que encontramos ejemplos donde algunas letras fueron permutadas, sólo analizaremos una, la *l*. Prescindiremos de las demás debido a que de cada una de ellas sólo obtuvimos un ejemplo y esto no nos ayuda para considerar patrones, en cambio, recolectando los pocos ejemplos de permutación de *l* encontramos pequeños patrones como se podrá ver:

- al cancha (*la*)
- la recreo (*al*)
- el sigimos (*le seguimos*)
- aPilcados (*aplicados*)
- Xochilt (*Xochitl*)
- blfores (*flores*)

En los primeros tres ejemplos observemos que hay una confusión entre el uso de los artículos y las preposiciones con artículo, así una vez más podemos notar que algo pasa alrededor de las preposiciones, por lo que sería interesante hacer un análisis del uso de éstas en textos escritos por niños, ya que eso puede ayudarnos a explicar lo que ocurre aquí y lo que ocurre con respecto a las omisiones de la preposición *a*.

Posteriormente vemos que los tres últimos ejemplos tienen más que ver con problemas fonológicos de distinta índole. En el caso de “aPilcados” podemos notar que la permutación de la *l* da como consecuencia una sílaba más fácil de pronunciar, a diferencia de “Xochilt”, en este caso tal parece que el niño no intentaba buscar la manera más sencilla de pronunciación, sino simplemente no reconoció cuál de las dos letras iba primero.

Finalmente con el último ejemplo pasa algo curioso. El niño sabe que hay un sonido compuesto /fl/ pero por alguna razón no puede relacionar ese sonido con las grafías *fl*, por lo que recurre a un conjunto de grafías que es más común encontrar en muchas palabras, el conjunto *bl*. Pero aunque usa este conjunto el niño sabe que éste no está caracterizado por un sonido fricativo, así que agrega la letra *f*.

6. CONCLUSIONES

A pesar de que el proyecto *Aprender mientras se enseña: una experiencia de acompañamiento en la enseñanza de la Lengua escrita* aun continúa, podemos estar seguros de que se ha alcanzado el objetivo de la tesis[46], debido a que ahora contamos con un corpus etiquetado de discurso infantil escrito que nos indica cuáles son los fenómenos ortográficas que se producen, la jerarquía numérica con la que se presentan y la frecuencia de uso sobre puntuación, mayúsculas, etc. Asimismo, aunque aún no se ha subido a la red, ya se cuenta con el diseño del corpus que será presentado a los usuarios, el cual es la hoja de estilo que describimos anteriormente. Por tanto, es claro que el trabajo aquí expuesto nos permite reflexionar sobre varias conclusiones que expondremos en este capítulo. Además, haremos algunas sugerencias para otros trabajos a futuro, ya que como hemos dicho, en esta tesis sólo exponemos el análisis que se hizo de algunos fenómenos que se presentan, pero el corpus puede proporcionar mucha más información.

Al analizar la primera gráfica comparativa[47] nos dimos cuenta que en los fenómenos más recurrentes el conjunto N tiene el mayor número de porcentaje, lo cual significa que los niños pertenecientes a escuelas donde los maestros no fueron capacitados son más propensos a cometer fenómenos ortográficos.

Asimismo, al hacerse la suma de todos los resultados que nos proporcionaron las tablas de resumen, también se hizo la suma de las palabras, lo cual nos dio por resultado que, en total, el conjunto N hizo 7,077 palabras y el conjunto S hizo 11,073 palabras, por lo que, según lo antes mencionado, los niños pertenecientes a las escuelas donde los maestros sí fueron capacitados son quienes tienen un desenvolvimiento narrativo. Por supuesto hay sus excepciones en ambos conjuntos, debido a que así como se pueden encontrar tres o cuatro textos muy bien escritos en el conjunto N, también se pueden encontrar algunos textos con demasiados fenómenos ortográficos en el conjunto S, pero en general, al ver los textos de ambos conjuntos sí es muy evidente la diferencia y los resultados expuestos en el capítulo anterior la hacen más notoria.

---

[46] Que como ya hemos dicho es uno de los objetivos del proyecto *Aprender mientras se enseña: una experiencia de acompañamiento en la enseñanza de la Lengua escrita*.

[47] Gráfica 1, pág. 84.

6.1. Algunas sugerencias para corregir el esquema

Como se recordará, para realizar un etiquetado en XML es necesario diseñar primero un esquema/DTD donde se especifique las etiquetas que se usarán y sus criterios de uso. Asimismo, contamos con el esquema XSLT Stylesheet, que a diferencia del otro no ayuda a etiquetar, sino que ayuda a diseñar la hoja de estilo, es decir, la presentación final de nuestro trabajo.

Ahora bien, el esquema/DTD y el esquema XSLT Stylesheet diseñados por el Mtro. Méndez para este corpus nos ayudaron a obtener los resultados anteriormente expuestos, los cuales por supuesto aportan información muy importante sobre los fenómenos que ocurren en los textos escritos por niños. Pero mientras realicé el análisis de los resultados me di cuenta que algunas cosas no se concretizaron, por lo que en este punto haré algunas sugerencias para modificar los esquemas y así obtener resultados que nos proporcionen información aún más detallada.

Antes de empezar es importante mencionar que estas sugerencias son sólo eso, no son modificaciones irrefutables debido a que éstas son consecuencia de la observación de sólo una persona, su servidora. Además, todo el trabajo para obtener este corpus se hizo de manera colectiva, por lo que las ideas que expondré las dejo a juicio de todos mis colegas para que se tomen las decisiones más adecuadas al respecto.

Ahora bien, si recordamos las etiquetas respecto a la puntuación podemos reconocer que esta etiqueta <r c="Fps"/> se utilizó para señalar los puntos que sirven únicamente para separar las palabras, pero a pesar de que lo marcamos en el etiquetado, la hoja de estilo no hace un conteo automático de este fenómeno. Como consecuencia omitimos el análisis de estos puntos, los cuales podrían ser considerados como segmentos, y si prescindimos de su análisis no obtendremos resultados confiables sobre la hipo y la hipersegmentación.

Continuando con los segmentos y tomando en cuenta el trabajo realizado por Emilia Ferreiro y colegas en *Caperucita roja aprende a escribir*, podemos considerar agregar atributos a las etiquetas de hipo e hipersegmentación. Para la etiqueta de hiposegmentación se requerirían dos tipos de atributo, uno que especifique el tipo de palabra [48] que antecede la hiposegmentación, y otro que especifique la palabra que la sigue. Con respecto a la hipersegmentación se necesitaría sólo un atributo que

[48] Artículo, preposición, verbo, etc.

especifique el tipo de palabra en el que se encuentra este fenómeno. Con esto, al igual que Ferreiro, no sólo obtendremos el porcentaje de estos fenómenos en los textos, sino que además podremos encontrar un patrón que nos ayude a entender cuáles son las causas por las que los niños pegan las palabras o segmentan algunas de ellas.

A pesar de que en nuestro objeto de estudio sólo encontramos tres ejemplos de agrandamiento de minúsculas, sería interesante agregar a la etiqueta, correspondiente a este fenómeno, un atributo donde se aclare si el niño agrandó una minúscula para darle uso de mayúscula. Esto nos ayudaría a entender mejor la existencia de estas grafías, y en caso de que tuvieran funcionamiento de mayúsculas podríamos hacer un mejor análisis del uso de éstas.

Asimismo, aunque tenemos una etiqueta para todas las mayúsculas iniciales que puso el niño, no tenemos una etiqueta que marque las mayúsculas iniciales que faltan. Podría considerarse que éstas últimas están contadas en las etiquetas de sustitución de mayúsculas por minúsculas, pero no todas estas sustituciones nos están señalando la exclusión de las mayúsculas iniciales, sino también la sustitución de mayúsculas en siglas. Así tenemos el número de mayúsculas iniciales que escribieron los niños, pero no tenemos la contraparte con la cuál hacer una comparación. Lo mismo ocurre con los acentos, sabemos cuántos acentos fueron omitidos, pero no podemos hacer un análisis de ellos si no sabemos cuántos acentos están correctamente usados.

Con respecto a las autocorrecciones he mencionado que el niño corrige lo que escribe con la o las grafías que cree que son las correctas, pero algunas veces en realidad lo que él cree no es lo convencional. Así, considero importante agregar a esta etiqueta un atributo donde se especifique si la corrección del niño entra dentro de lo convencional, ya que aunque en el etiquetado se complementa la información con otra etiqueta cuando la autocorrección del niño no es lo normativo (como se ve en el ejemplo de abajo), en la tabla de resumen no se especifica esa complementación de información.

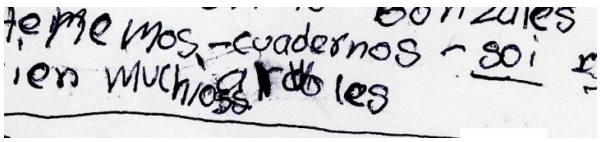
	<g n="árboles"/> ar <v>oles <corrección tipo="susti"/> <sus tipo= "bxv"/>

Ilustración U Autocorrecciones

Como podemos ver, el niño sustituyó una *d* por una *v*, pero aún con esa sustitución la palabra no está correctamente escrita, por lo que se usó la etiqueta <sus/> para señalar la sustitución de la grafía *b*. Así, aunque en el etiquetado vemos esta complementación de información, en la tabla de resumen no la vemos, lo único que encontramos es la suma de todas las correcciones, tanto de las que están bien como de las que no. Esta información nos ha permitido tener un panorama general de la frecuencia con que los niños reflexionan sobre lo que escriben, pero me parece de vital importancia reconocer las autocorrecciones que están bien de las que están mal, ya que si estamos analizando los fenómenos ortográficos es importante saber todo lo que se pueda sobre ellas. Además, desde mi parecer, no tiene el mismo valor una transgresión que el niño hizo inconscientemente que una que hizo incluso después de haber reflexionado.

De la misma forma creo que es importante agregar un atributo a las etiquetas de puntuación y <l tipo="conguion"/> donde se especifique si el uso de éstas es correcto.

Con respecto a la puntuación podemos encontrar signos que no tienen el valor que la norma establece, como en el ejemplo "mi escuela es . grande". En todos los ejemplos semejantes que se encontraron se dejó la etiqueta de punto sin especificar si era seguido, de abreviatura, etc. ya que no pertenece a ninguno de éstos, pero en la tabla de resumen no aparece el conteo de todas estas etiquetas indeterminadas.

De esta forma sugiero tres ideas para resolver esto: a) agregar un atributo que especifique cuando un punto se usa de forma correcta; b) agregar un atributo que especifique cuando un punto se usa arbitrariamente; o c) configurar el esquema XSLT

Stylesheet para que haga un conteo de las etiquetas de punto que se quedan sin especificar.

Considero importante diferenciar los puntos que parecen tener un uso normativo de los que no lo parecen, porque dependiendo de la cantidad de estos últimos podemos concluir si realmente el niño sabe usar o no los puntos. Es como vimos con el uso de las mayúsculas: al notar que los niños del conjunto N hacían un uso excesivo de mayúsculas poníamos en tela de juicio si las mayúsculas que parecían estar bien eran arbitrarias o no. Lo mismo pasa con los puntos, si hay una gran cantidad de puntos distribuidos en todo el texto podemos dudar de la validez de los puntos que parecen estar bien.

En relación con la etiqueta <l tipo="conguion"/> creo que es importante diferenciar si el niño segmentó bien o mal la palabra, ya que de ahí podríamos analizar la forma en cómo separan las sílabas, lo cual podría complementar el análisis de la segmentación.

Finalmente sólo me quedan algunas sugerencias respecto al diseño de nuestra hoja de estilo:

Tomando en cuenta el trabajo del INEE se podría configurar nuestra hoja de estilo para que el usuario pueda extraer los textos del corpus de forma selectiva, ya sea de acuerdo con el género, edad, conjunto, zona, sector, etc.

También sería útil que el usuario pudiera extraer sólo los fenómenos que necesita analizar, por ejemplo, si requiere ver las omisiones podría abrirse una ventana con una lista de todas las palabras con omisiones. Asimismo, en este listado se podrían crear ligas que lleven al usuario a la imagen del texto donde se encuentra el fenómeno que está analizando.

Por último, ayudaría mucho crear hipervínculos en las palabras que tienen pico-paréntesis para ir directamente a su comentario correspondiente, y viceversa, del comentario ir al ejemplo.

6.2. Trabajo futuro

Con todas las sugerencias que propuse en el punto anterior es de esperarse que aún haya mucho trabajo por hacer. Pero como bien se ha mencionado antes, el corpus de este proyecto puede ser utilizado para muchos otros análisis. Algunos de ellos pueden ser a nivel morfológico, sintáctico, discursivo, etc. Y aún quedarían muchas otras cosas por

analizar, por ello en este punto propondré algunos trabajos de investigación que, con los permisos correspondientes, podrían servirse del corpus y del material del proyecto *Aprender mientras se enseña: una experiencia de acompañamiento en la enseñanza de la Lengua escrita*.

Creo que sería interesante hacer un trabajo similar al que se hizo en esta tesis, sólo que además de recolectar archivos escritos por niños, podríamos recolectar archivos orales de los mismos niños. Así, además de hacer un análisis de la escritura se puede hacer a la par un análisis de la oralidad e ir comparando los fenómenos que ocurren en ambos discursos. Este trabajo nos permitiría descubrir qué tan apegado es el discurso escrito al discurso oral en niños. Asimismo, podría considerarse como un trabajo dialectal infantil de la zona donde se recolecte.

Con base en el panorama general que abrió este proyecto, se pueden hacer varios análisis de elementos en particular que se encuentran en el corpus, por ejemplo, se podría hacer un estudio únicamente del uso de las preposiciones, ya que como vimos en los resultados, algunas de las omisiones de *a* coincidían con que ésta tenía función de preposición, por tanto, además de conocer el uso que le dan los niños a las preposiciones, un análisis sobre estos elementos nos ayudaría a complementar la información que se ha expuesto.

También se puede hacer un análisis de los verbos *ser* y *estar*, ya que durante la normalización noté algunos casos donde parece que los niños no tienen claro cuándo usar tal o cual verbo, y en consecuencia usaban el verbo *ser* cuando debieron usar el verbo *estar*, o viceversa.

Asimismo, se puede explorar la conjugación de los verbos, debido a que hay muchos casos donde no existe una concordancia entre éstos y el sujeto. Con esta exploración se podría saber qué verbos son los más difíciles para los niños y qué estructuras sintácticas pueden estar involucradas en la falta de concordancia.

En relación con la concordancia se encontraron muchos casos en los que no hay correspondencia de número y género entre los elementos de una frase nominal, por tanto, sería interesante analizar estas frases para encontrar posibles causas de este fenómeno.

Otro fenómeno que me parece interesante estudiar a fondo son las autocorrecciones. Es interesante comprender las reflexiones que hace el niño sobre lo que está escribiendo, debido a que esto también nos permite observar la manera en la que concretiza el discurso escrito. Así, analizar los elementos que tacha y lo que dice

alrededor de ellos, los elementos que pone sobre otros, etc., podría decirnos mucho sobre lo que sabe el niño y lo que cree que es correcto en la escritura. La gran mayoría de los estudios sobre el aprendizaje de la escritura en niños y adultos toman muy en cuenta los errores que cometen, pero hasta ahora no ha habido estudios que presten atención a las autocorrecciones, por tanto, supongo que analizar este fenómeno puede aportar información innovadora sobre la adquisición del discurso escrito.

Estas son mis propuestas para trabajos a futuro, por supuesto algunos que se interesen pueden encontrar en éstas objetivos más convincentes o resultados más satisfactorios que los que menciono, o también pueden encontrar otros temas de análisis que ofrece el corpus del proyecto.

BIBLIOGRAFÍA

Libros y revistas

Abaitua, Joseba. «Tratamiento de corpora bilingües. » Tratamiento del lenguaje natural (2002): 61-90.

Aguilar, César Antonio. *¿Mecuentas un cuento?*: Relaciones entre frases nominales, referencia y cohesión en narraciones orales infantiles. Universidad Nacional Autónoma de México. Facultad de Filosofía y Letras. Posgrado en Lingüística. 2003.

Alarcos Llorach, Emilio. Fonología española. Madrid: Gredos, 1965.

Albaladejo, Tomás. «Retórica, tecnologías, receptores.» Revista de Retórica y Teoría de la Comunicación (2001): 9-18.

Albeentosa Hernández, José Ignacio. Moya Guijarro, Arsenio Jesús. Narración infantil y discurso (Estudio lingüístico de cuentos en castellano e inglés). Cuenca: Universidad de Castilla - La Mancha, 2001.

Arrearte, Gerardo. «Normas y estándares para la codificación de textos y para la ingeniería lingüística. » Filología e informática: Nuevas tecnologías en los estudios filológicos. (1999): 17- 44.

Barriga Villanueva, Rebeca. Estudios sobre habla infantil en los años escolares "...un solecito calientote...". México: Colegio de México, 2002.

Blanche Benveniste, Claire. «La escritura, irreductible a un "código"». Relaciones de (in) dependencia entre oralidad y escritura. comp. Emilia Ferreiro. Barcelona: Gedisa, 2002.

Bolaño e Isla, Amancio. Breve manual de fonética elemental: sonidos correctos e incorrectos del español de México. México: Porrúa, 1968.

Cantero, Francisco José. Arriba, José de. Psicolingüística del discurso. Barcelona: Octaedro, 1997.

Downing, John. «La influencia en la escuela en el aprendizaje de la lectura». Nuevas perspectivas sobre los procesos de lectura y escritura. comp. Emilia Ferreiro. México: Siglo XXI, 1982.

Fernando Lara, Luis. «La escritura como tradición y como instrumento de reflexión». Relaciones de (in) dependencia entre oralidad y escritura. comp. Emilia Ferreiro. Barcelona: Gedisa, 2002.

Ferreiro, Emilia. Pontecorvo, Clotilde. Et al. Caperucita roja aprende a escribir. Estudios psicolingüísticos comparativos en tres lenguas. Barcelona: Gedisa, 1996.

Ferreiro, Emilia. Teberosky, Ana. Los sistemas de escritura en el desarrollo del niño. México: Siglo XXI, 1982.

Ferreiro, Emilia. Nuevas perspectivas sobre los procesos de lectura y escritura. México: Siglo XXI, 1982.

Ferreiro, Emilia. Relaciones de (in) dependencia entre oralidad y escritura. Barcelona: Gedisa, 2002.

Galeote Moreno, Miguel. Adquisición del lenguaje. Problemas, investigación y perspectivas. Madrid: Pirámide, 2002.

Garvey, Catherine. Children's talk. Trad. Alfredo Guera Miralles. 1. Madrid: Morata, 1984.

Gee, James Paul. An introduction to discourse analysis (Theory and method). New York: Routledge, 2005.

Garside, R., Leech, G., McEnery, A. Corpus Annotation: Linguistic Information from Computer Text Corpora. New York: Addison Wesley Longman, 1997.

Gil Fernández, Juana. Fonética para profesores de español: de la teoría a la práctica. Madrid: arco/ libros, 2007.

González Sánchez, Margarita. Lenguaje escolar y clase social. Salamanca: Amarú, s.d.

Goldfarb, Charles. Prescod, Paul. Manual de XML. Madrid: Pentrice Hall Iberia, 1999.

Halliday, M. A. K. The language of early childhood. Vol. 4. New York: Continuum, 2004. 10 vol.

Harris, Roy. Signos de escritura. Barcelona: Gedisa, 1999.

Introno, Francesco di. Teso, Enrique del. Et al. Fonética y fonología actual del español. Madrid: Catedra, 1995.

Iribarren, Mary C. Fonética y fonología españolas. Madrid: Síntesis, 2005.

Leal García, Aurora. Construcción de sistemas simbólicos: la lengua escrita como creación. Barcelona: Gedisa, 1987.

Maingueneau, D. Introducción a los métodos de análisis del discurso. Buenos Aires: Hachette, 1976.

Moreno Boronat, Lidia. Palomar Sanz, Manuel. Et al. Introducción al procesamiento de lenguaje natural. Murcia: Universidad de Alicante, 1999.

Oesterreicher, Wulf. «Pragmática del discurso oral». Berg, W.B. Oralidad y argentividad. Tübingen: Narr, 1996.

Olson, Davis R. Torrance, Nancy. Cultura escrita y oralidad. Barcelona: Gedisa, 1995.

Pardo, Neyla. Cómo hacer análisis crítico del discurso (una perspectiva latinoamericana). Santiago de Chile: Frasis, 2007.

Pérez Grajales, Héctor. Comunicación escrita. Producción e interpretación del discurso escrito (Talleres). Bogotá: Magisterio, 1995.

Pérez Guerra, Javier. Introducción a la lingüística de corpus. Un ejercicio con herramientas informáticas aplicadas al análisi textual. Santiago de Compostela: Torculo Edición, 1998.

Pontecorvo, Clotilde. «Las prácticas de alfabetización escolar: ¿es aún válido el "hablar bien para escribir bien"?». Relaciones de (in) dependencia entre oralidad y escritura. comp. Emilia Ferreiro. Barcelona: Gedisa, 2002.

Quilis Antonio. Tratado de fonología y fonética españolas. Madrid: Gredos, 1999.

Sierra, Gerardo. «Diseño de corpus textuales para fines lingüísticos». IX Encuentro Internacional de Lingüística en el Noroeste. Tomo II (2008): 445- 462.

Sierra, Gerardo. Rosas, Alejandro. «Propuesta de clasificación para corpus lingüísticos informatizados». X Encuentro Internacional de Lingüística en el Noroeste. (en prensa).

Sinclair, Hermine. «El desarrollo de la escritura: avances, problemas y perspectivas». Nuevas perspectivas sobre los procesos de lectura y escritura. comp. Emilia Ferreiro. México: Siglo XXI, 1982.

Torruella, J. Llisterri, J. «Diseño de corpus textuales y orales». Filología e informática. Nuevas tecnologías en los estudios filológicos (1999): 45-77.

Internet

Google. ASCII table and description. 11 abril 2009. <http://www.asciitable.com/>

INEE. Cospus Excale de Escritura. 22 octubre 2009. < http://www.inee.edu.mx/index.php?option=com_wrapper&view=wrapper&Itemid =1318>

Ingeniería Lingüïstica. Lingüística de corpus. 5 abril 2010. <http://www.iling.unam.mx/CursoCorpus/default.html>

MacWhinney, Brian. CHILDES. 6 abril 2010. 18 octubre 2003. <http://childes.psy.cmu.edu/>

MacWhinney, Brian (coord). Talkbank. 30 marzo 2010. 07 noviembre 2003. <http://talkbank.org/>

Pérez Hernández, Chantal. «Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento. » Estudios de Lingüística del Español. (2002). http://elies.rediris.es/elies18/233.html

Silva V., Omer. El análisis del discurso según Van Dijik y los estudios de la comunicación. Abril-mayo 2002. 19 marzo 2009. <http://www.cem.itesm.mx/dacs/publicaciones/logos/anteriores/n26/osilva.html>