**ARTICULO ORIGINAL**

# DISEÑO DE UN CORPUS LINGÜÍSTICO

*DESIGN OF A LINGUISTIC CORPUS*

Ayala Nieto Ángela Patricia. [1]; Guerrero Quintero Natalia Andrea. [2]; Mogollón Tolosa Mabel Xiomara. [3]; Portilla Portilla Edwin Mauricio. [4]; Rangel Navia Heriberto José [5]., Contreras Velásquez Zaida Rocío [6].

## RESUMEN

**INTRODUCCIÓN:** El objetivo de la investigación fue diseñar un corpus lingüístico del lenguaje infantil en niños y niñas de 3 años 11 meses a 4 años 11 meses pertenecientes a Centros de Desarrollo Infantil en el municipio de Cúcuta. **MÉTODOS:** Se realizó un estudio cuantitativo descriptivo de corte transversal, cuyo universo de estudio estuvo constituido por 92 niños, se realizaron grabaciones durante 8 semanas 3 días a la semana, en videos de 20 minutos dentro del salón de clases. La muestra estudiada fueron 22 niños, a los cuales se les realizó transcripción por medio del software SALT. **RESULTADOS:** Corpus de tipología oral, que consta de 7,498 palabras, distribuidas en 22 textos, con un total de 4 horas y 20 minutos de grabación y 22 participantes (12 niñas, 10 niñas). **ANÁLISIS Y DISCUSIÓN:** La principal contribución de este corpus es la aportación de un recurso lingüístico que aún hoy escasea. Este corpus, facilita una muestra de la variedad lingüística infantil, tanto en formato de audio como de texto. Además, dichos textos están enriquecidos con codificaciones, de lo cual se puede extraer automáticamente todo tipo de información referente a los niveles de sintaxis, semántica, discurso, velocidad, fluidez, errores y omisiones. **CONCLUSIONES:** El corpus cumple con la característica de los corpus modernos de ser formato electrónico, permitiendo el almacenamiento y la manipulación de los datos y su posible intercambio con otros investigadores interesados, pero no cumple con la característica de representatividad, porque su diversidad en cuanto a los rangos de edad es baja.

**PALABRAS CLAVE:** Lingüística, corpus oral, SALT.

[1]. Fonoaudióloga, Docente de Universidad de Pamplona, Especialista en Práctica Pedagógica Universitaria.

[2]. Estudiante de fonoaudiología Universidad de Pamplona.

[3]. Fonoaudióloga, Docente de Universidad de Pamplona, Mg. Educación.

[4]. Fonoaudiólogo, Docente Universidad de Pamplona, Mg. Educación.

[5]. Fonoaudiólogo, Docente Universidad de Pamplona, Mg. Educación.

[6]. Bacterióloga Clínica, Docente Universidad de Pamplona, Especialista Epidemiología.

## ABSTRACT

**INTRODUCTION:** The objective of the research was to design a linguistic corpus of children's language in children from 3 years 11 months to 4 years 11 months belonging to Child Development Centers in the municipality of Cúcuta. **METHODS:** A cross-sectional descriptive quantitative study was carried out, whose universe of study consisted of 92 children. Recordings were made for 8 weeks 3 days a week, in videos of 20 minutes in the classroom. The sample studied was 22 children, who underwent transcription through SALT software. RESULTS: Corpus of oral typology, consisting of 7,498 words, distributed in 22 texts, with a total of 4 hours and 20 minutes of recording and 22 participants (12 girls, 10 girls). **ANALYSIS AND DISCUSSION:** The main contribution of this corpus is the contribution of a linguistic resource that is still scarce today. This corpus facilitates a sample of the children's linguistic variety, both in audio and text formats. In addition, these texts are enriched with encodings, from which all types of information can be extracted automatically referring to the levels of syntax, semantics, discourse, speed, fluency, errors and omissions. **CONCLUSIONS:** The corpus complies with the characteristic of modern corpora of being an electronic format, allowing the storage and manipulation of data and its possible exchange with other interested researchers, but does not comply with the characteristic of representativeness, because diversity in terms of ranges old is low.

**KEY WORDS:** Linguistics, oral corpus, SALT.

# INTRODUCCIÓN

La Lingüística de Corpus (LC) es una disciplina consolidada dentro de los estudios lingüísticos. En la actualidad, se define como el estudio del lenguaje a partir de corpus, entendiendo corpus como una gran colección de textos disponibles en formato electrónico. Se pueden distinguir dos tipos de corpus, corpus textuales y orales. Los corpus orales se clasifican en dos categorías: corpus para el estudio de la lengua oral y corpus para el desarrollo de aplicaciones en tecnologías de habla, el primero tiene como propósito principal caracterizar desde un punto de vista lingüístico la lengua hablada (1).

De acuerdo a lo anterior, el objetivo de la investigación fue diseñar un corpus lingüístico del lenguaje infantil en niños y niñas de 3 años 11 meses a 4 años 11 meses pertenecientes a Centros de Desarrollo Infantil en la ciudad de San José de Cúcuta y se abordó, a partir, de la pregunta de investigación ¿Cuáles son los umbrales de desarrollo del lenguaje oral en infantes? Es por esto, que la interacción interpersonal es el espacio donde el lenguaje se hace consensual, operando en la dinámica interaccional principios y reglas que permiten comunicarse. Los individuos se comunican con reglas, siendo el conocimiento de las reglas lingüísticas y la capacidad de aplicarlas en determinados contextos, lo que constituye la base de la competencia comunicativa (2).

El contexto social desempeña un rol fundamental en el aprendizaje de las locuciones tempranas, proporcionando la estructura y contenido de éstas. Los factores situacionales, que a veces condicionan qué es lo que puede decir el niño, incluyen objetos, actividades y personas de la acción comunicativa, así como otras variables conversacionales. También, los factores internos del niño, que incluyen su percepción de la situación, que influyen en las primeras locuciones. En esta etapa inicial del desarrollo lingüístico el niño comunica más de lo que puede codificar, lo cual se demuestra en la capacidad de los adultos de adivinar el mensaje que pretende enviar el niño (3).

Las relaciones de roles sociales varían con los diferentes contextos. Los niños en algunos contextos asumen un rol de importancia central, y así son capaces de jugar un rol dominante en muchas de sus actividades diarias, partiendo de lo anterior, podría decirse que el contexto se "centra en la interacción del niño", Mientras las relaciones entre roles varían con el contexto, el evento y las relaciones de poder entre participantes, esto se analiza, a través, de aspectos comunes para algunos niños en los roles que están asignados sin considerar el contexto (4).

Los alcances de la investigación se enfocaron en establecer umbrales del lenguaje oral de la población estudiada, a través del análisis de frecuencia de los datos obtenidos por medio del Software SALT, que permitió el estudio en los niveles de sintaxis, semántica y discurso de los oradores.

La pretensión del estudio, se fundamentó en el diseño del Corpus Lingüístico, para posterior divulgación del mismo, que permitirá la codificación escrita, muestras y comportamientos de fluidez, generando así, una base de datos de referencia, con el análisis detallado de los niveles de lenguaje del habla infantil.

## MÉTODOS

Se realizó un estudio cuantitativo descriptivo de corte transversal, cuyo universo de estudio estuvo constituido por 92 niños, pertenecientes a Centros de Desarrollo Infantil del municipio de San José de Cúcuta, de edades comprendidas entre 3 años 11 meses y 4 años 11 meses.

Se definió, la muestra del estudio en tres fases; en la primera fase se grabaron 92 niños durante 8 semanas 3 días a la semana, en videos de 20 minutos con 2 cámaras de calidad semi-profesional con sus respectivos trípodes, 2 micrófonos bidireccionales, los cuales grababan de manera simultánea grupos de 2 a 4 niños que se ubicaban en mesas dentro del salón de clases, acompañados siempre de un examinador.

Una segunda fase, que consto en la revisión de la calidad de audio e imagen de la totalidad de los videos, para así determinar cuáles videos eran viables para transcripción. En esta fase fueron eliminadas las muestras de video de 14 niños no aptas debido a condiciones técnicas no favorables como: fallas en el audio, falla de origen de la memoria SD, quedando en esta fase 78 vídeos disponibles para el estudio.

Y por último, la tercera fase, cada toma audiovisual pasó por cuatro criterios de inclusión establecidos en el libro de referencia “EVALUACIÓN DE LA PRODUCCIÓN DE LENGUAJE MEDIANTE EL SOFTWARE DE SALT: Guía clínica para el análisis de muestras de lenguaje (2ª edición)” disponible en la página web http://saltsoftware.com/, los cuales determinan una muestra de lenguaje válida para transcripción:

1. El examinador sigue el protocolo de obtención.
2. El protocolo de obtención desafía las habilidades de producción del hablante.
3. El hablante produce una muestra que es representativa en su idioma.
4. Al menos el 80% de la muestra es inteligible.

De acuerdo a lo anterior, con el primer criterio se eliminaron 17 muestras de vídeo ya que no cumplían con el protocolo de conversación, el cual fue escogido para esta investigación. En el segundo y tercer criterio fueron eliminados 26 videos, y con el último criterio se descartaron 13. Por consiguiente fueron 22 tomas de vídeo la muestra de estudio para esta investigación.

Después de haber definido la muestra se inició el proceso de transcripción, estas se hicieron con el software Análisis Sistemático de Transcripciones de Lenguaje (SALT) 2012 Clinical Software, con licencia 099-34482-41215, sistema de operación windows. SALT es un software que estandariza el proceso de obtención, transcripción y análisis de muestras de lenguaje. Este proporciona informes de análisis instantáneos a partir de medidas estándar de sintaxis, semántica, discurso, velocidad, fluidez, errores y omisiones.

El software cuenta con bases de datos de más de 7,000 oradores típicos. Los participantes varían en edad, sexo, nivel socioeconómico estado y ubicación geográfica, se usaron diferentes protocolos de licitación para recoger las muestras, y cada muestra incluida en una base de datos fue provocada siguiendo el protocolo correspondiente. SALT permite comparar de manera automática la muestra objetivo con un grupo de edad o grado escolar similar de las bases de datos disponibles en la plataforma (5).

Cabe resaltar que antes de iniciar las grabaciones, la totalidad de los acudientes de los menores autorizaron el proceso través de un consentimiento informado donde se especificaba que estos videos serían realizados con fines investigativos.

Los datos arrojados por el software SALT fueron ingresados y organizados en una base de datos diseñada especialmente para el estudio en el programa Excel.

El análisis de la información obtenida en el presente estudio fue llevada a cabo en el software estadístico SPSS versión 21, este se usó para la comparación de medias a través de la prueba T de student o Test-T, esta prueba permite comparar muestras y/o establece la diferencia entre las medias de las muestras (6), es así que se compararon las medias de los valores hallados en esta investigación con las medias de los datos de una base de datos disponible en el software SALT con edades y número de participantes similares. Además se establecieron los puntos de corte, que son hallados a partir de la mediana que es una medida de tendencia central, y análisis de frecuencias que permitió obtener una descripción de la distribución de las variables.

# RESULTADOS

Este corpus es de tipología oral, formado por las transcripciones realizadas a partir de grabaciones de habla espontánea, consta de 7,498 palabras, distribuidas en 22 textos, con un total de 4 horas y 20 minutos de grabación y 22 participantes (12 niñas, 10 niñas). Las transcripciones se encuentran en tipo de archivo SALT. Transcript (.slt) y los vídeos en formato MP3. Además de los archivos de audio y texto, se realizó la base de datos en formato Excel que contiene datos personales y los valores arrojados por el software SALT de las variables de los niveles de sintaxis, semántica, discurso, velocidad, fluidez, errores y omisiones.

Con relación al análisis estadístico se realizara a continuación la descripción. En primer lugar se hizo la comparación de medias mediante el Test-T usando el software SPSS, comparando los datos obtenidos en esta investigación con los de una base de datos proporcionada por SALT, como se muestra a continuación:

| VARIABLE | MEDIA (Investigación) | MEDIA (Base de datos) | P | IC 95% |
|---|---|---|---|---|
| **LONGITUD DE TRANSCRIPCION** | | | | |
| Total enunciado | 68,91 | 96,5 | 0,000 | -38,95 - (-16,23) |
| Conjunto de análisis | 58,86 | 90,77 | 0,000 | -41,90 - (-19,91) |
| Total palabras completadas | 340,82 | 431,88 | 0,020 | -166,28 - (-15,84) |
| **SINTAXIS/MORFOLOGIA** | | | | |
| MLU en palabras | 5,17 | 4,23 | 0,006 | 0,30 - 1,58 |
| MLU en morfemas | 5,27 | 4,65 | 0.066 | -0,05 – (-1,29) |
| **SEMANTICA** | | | | |
| Número diferente de palabras | 135,00 | 137,04 | 0,851 | -24,31 – (-20,23) |
| Número total de palabras | 310,41 | 368,73 | 0,079 | -124,12 – 7,48 |
| Tipo Token Ratio | 0,46 | 0,4 | 0,001 | 0,03 – 0,09 |
| **DISCURSO** | | | | |
| % Respuestas a preguntas | 91,59 | 71,01 | 0,000 | 17,18 – 23,98 |
| Longitud media de vuelta | 6,19 | 5,99 | 0,711 | -0,92 – 1,32 |
| Enunciados discurso superación | 1,23 | 11 | 0,000 | -10,80 – (-8,75) |
| Interrupciones de otro orador | 0,55 | 1,5 | 0,000 | -1,38 – (-0,53) |
| **INTELIGIBILIDAD** | | | | |
| % Inteligibilidad enunciados | 96,73 | 95,95 | 0,364 | -0,97 – 2,52 |
| **LABERINTOS Y ABANDONOS DE ENUNCIADOS** | | | | |
| Enunciados con laberintos | 2,55 | 20,54 | 0,000 | -19,39 – (-16,60) |
| Número de laberintos | 3,27 | 24,88 | 0,000 | -23,36 - (-19,85) |

| | | | | |
|---|---|---|---|---|
| Número de laberintos de palabras | 4,50 | 51,58 | 0,000 | -49,49 - (44,67) |
| Laberintos palabras como % palabras totales | 1,23 | 12,26 | 0,000 | -11,56 – (-10,50) |
| Enunciados abandonados | 0,64 | 1,65 | 0,000 | -1,34 – (-0,69) |
| **FACILIDAD VERBAL Y RITMO** | | | | |
| Palabras/minuto | 16,55 | 52,08 | 0,000 | -38,70 – (-32,36) |
| Pausas dentro de enunciados | 0,64 | 1,58 | 0,000 | -1,39 – (-0,50) |
| Tiempo de pausas dentro de enunciado | 0,3 | 0,08 | 0,000 | -0,07 – (-0,02) |
| **OMISIONES Y CODIGO DE ERROR** | | | | |
| Palabras omitidas | 0,5 | 2,35 | 0,000 | -2,40 – (-2,21) |
| Morfemas omitidos | 0,5 | 1,38 | 0,000 | -1,43 – (-1,24) |
| Clíticos omitidos | - | - | - | - |
| Error código a nivel de palabra | 0,23 | 4,58 | 0,000 | -4,74 – (-3,97) |
| Error código a nivel de enunciado | 1,36 | 3,35 | 0,000 | -2,85 – (-1,13) |

TABLA 1. Comparación de medias, datos de presente estudio vs base de datos SALT. Fuente: los autores.

La tabla 1 muestra la comparación de medias usando el método de análisis estadístico con distribución por Test T. Las probabilidades de 0,05 o inferiores se consideran significativas, en los intervalos de confianza del 95%. No puede calcularse la variable clíticos omitidos porque la desviación típica es 0.

Seguidamente a partir de la mediana de los valores se determinaron los puntos de corte, como se muestra en la tabla 2.

| VARIABLE | BAJO | MEDIO | ALTO | RI |
|---|---|---|---|---|
| **LONGITUD DE TRANSCRIPCION** | | | | |
| Total enunciado | 51,50 | 63,00 | 79,25 | 27,7500 |
| Conjunto de análisis | 42,00 | 53,50 | 72,75 | 30,7500 |
| Total palabras completadas | 225,75 | 331,00 | 412,00 | 186,2500 |
| **SINTAXIS/ MORFOLOGIA** | | | | |
| MLU en palabras | 3,98 | 5,46 | 6,38 | 2,4000 |
| MLU en morfemas | 4,00 | 5,50 | 6,47 | 2,4675 |
| **SEMANTICA** | | | | |
| Número diferente de palabras | 97,25 | 132,00 | 167,75 | 70,5000 |
| Número total de palabras | 208,25 | 302,50 | 374,25 | 166,0000 |
| Tipo Token Ratio | 0,41 | 0,47 | 0,51 | 0,1050 |
| **DISCURSO** | | | | |
| % Respuestas a preguntas | 87,50 | 94,00 | 97,25 | 9,7500 |
| Longitud media de vuelta | 4,66 | 5,92 | 8,00 | 3,3425 |

| | | | | |
|---|---|---|---|---|
| Enunciados discurso superación | 0,00 | 0,00 | 1,25 | 1,2500 |
| Interrupciones de otro orador | 0,00 | 0,00 | 1,00 | 1,0000 |
| **INTELIGIBILIDAD** | | | | |
| % Inteligibilidad enunciados | 94,75 | 98,00 | 100,00 | 5,2500 |
| **LABERINTOS Y ABANDONOS DE ENUNCIADOS** | | | | |
| Enunciados con laberintos | 0,00 | 2,00 | 3,00 | 3,0000 |
| Número de laberintos | 0,00 | 2,00 | 4,25 | 4,2500 |
| Número de laberintos de palabras | 0,00 | 3,50 | 6,00 | 6,0000 |
| Laberintos palabras como % palabras totales | 0,00 | 1,00 | 2,00 | 2,0000 |
| Enunciados abandonados | 0,00 | 0,50 | 1,00 | 1,0000 |
| **FACILIDAD VERBAL Y RITMO** | | | | |
| Palabras/minuto | 11,40 | 16,55 | 20,60 | 9,2000 |
| Pausas dentro de enunciados | 0,00 | 0,00 | 1,00 | 1,0000 |
| Tiempo de pausas dentro de enunciado | 0,00 | 0,00 | 0,07 | 0,0725 |
| **OMISIONES Y CODIGO DE ERROR** | | | | |
| Palabras omitidas | 0,00 | 0,00 | 0,07 | 0,0000 |
| Morfemas omitidos | 0,00 | 0,00 | 0,00 | 0,0000 |
| Clíticos omitidos | 0,00 | 0,00 | 0,00 | 0,0000 |
| Error código a nivel de palabra | 0,00 | 0,00 | 0,00 | 0,0000 |
| Error código a nivel de enunciado | 0,00 | 1,00 | 2,00 | 2,0000 |

TABLA 2. Propuesta puntos de corte de las variables de investigación. Fuente: los autores.

La tabla 2 expone los puntos de corte de todas las variables, en las que se establecieron los cuartiles por bajo, medio y alto, por cuartiles, siendo así el parámetro a establecer la distribución de la muestra de estudio.

Para finalizar la descripción de resultados, el análisis de frecuencias en la figura 1 expone la distribución de las variables analizadas, a través de un histograma.

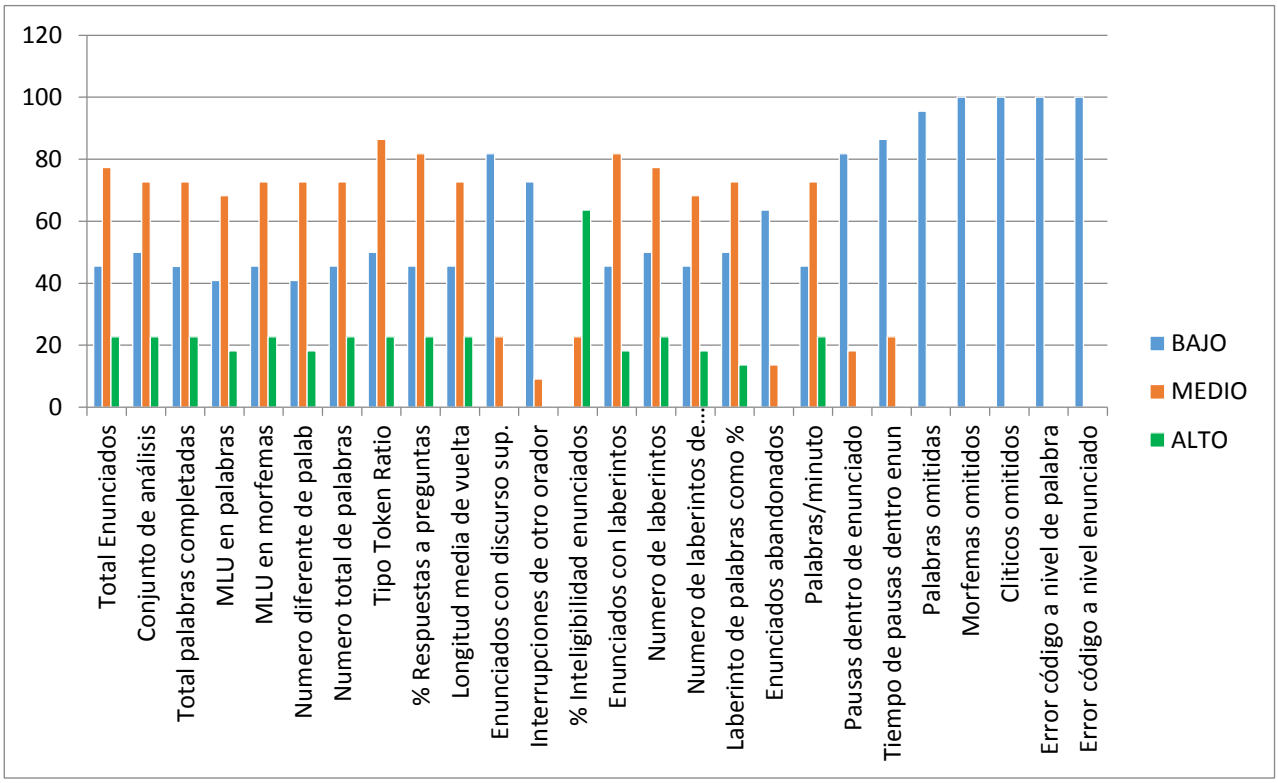


FIGURA 1. Análisis de frecuencias de las variables estudiadas. Fuente: los autores.

La figura 1 representa un análisis en el que se evalúa las frecuencias encontradas en el estudio correspondientes a las variables establecidas por el software SALT, en el que se determina en porcentaje, con un rango en forma de intervalo [0, 100].

## ANÁLISIS Y DISCUSIÓN

La mayor parte del trabajo realizado en LC se ha centrado en la lengua inglesa. El primer corpus moderno fue el Brown Corpus en lengua inglesa, y a partir de ese momento los recursos lingüísticos incrementaron en dicho idioma. De la misma manera, muchos países europeos y asiáticos siguieron el modelo y crearon sus corpus de referencia para sus propias lenguas (7).

En España los primeros trabajos en LC empezaron a llevarse a cabo en la década de los noventa. Así, podemos hablar de CORLEC (Corpus Oral de Referencia de la Lengua Española Contemporánea) como proyecto pionero. Fue el primer corpus de habla espontánea del español, bajo la dirección de F. Marcos Marín, recogido entre 1991 y 1992 y financiado por IBM. También se destaca el corpus C-ORAL-ROM, recopilado, entre otros equipos europeos, por el Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid. Se trata de un corpus oral multilingüe y general del habla de cuatro lenguas romances (francés, español, italiano y portugués) en sus diferentes registros (8).

A diferencia de países de lengua inglesa, asiáticos y europeos, los países latinoamericanos están en desventaja en estudios de LC, es hasta en esta última década que se han desarrollado corpus de lenguaje oral en esta región. El Corpus oral del español de México (COEM) y el Corpus Lingüísticos del Instituto Caro y Cuervo (CLICC) en Colombia son muestra de ello, estos dos son corpus del habla adulta. En Latinoamérica es escaso por no decir casi nulo los estudios de LC en habla infantil (8).

La necesidad de contar con datos reales para el análisis lingüístico ha supuesto un importante impulso en la construcción de corpus, convertidos en la actualidad en la base de la descripción, explicación y teorización lingüística en cualquiera de sus múltiples perspectivas o enfoques (9).

Elaborar un corpus constituye un trabajo complejo. Las labores de recopilación, transcripción y almacenaje van unidas a lo que se entiende por un corpus lingüístico, corpus que, además, actualmente debe reunir una serie de condiciones, como el hecho de ser electrónico, representativo, reproducible, y, ante todo, útil (10).

El formato del corpus diseñado en esta investigación es electrónico, permitiendo el almacenamiento y la manipulación de los datos desde cualquier ordenador. Partiendo del principio de que todo corpus es finito no se puede aspirar a contener todos los datos de una lengua. La finitud de un corpus obliga a que el diseño del mismo deba realizarse pensando más en la proporción y en la diversidad que en su tamaño (11). Por tanto, este corpus no cumple con la condición de representatividad, debido a la poca diversidad en el rango de edad de la muestra estudiada.

Según la clasificación establecida por la Text Encoding Iniciative (TEI), el corpus creado es de tipología especial. El corpus especial se centra sólo en el estudio del lenguaje de un grupo social, de una zona, etc. Suele ser pequeño y no debe cumplir la condición de representatividad (7).

La ventaja de los corpus modernos es su carácter digital y su tratamiento y análisis mediante el uso del ordenador. Los softwares disponibles en la actualidad facilitan enormemente el trabajo y ahorran mucho tiempo. SALT permite realizar tareas como la recuperación de texto herramientas de transcripción para facilitar el proceso, informes de análisis instantáneos, comparación de la muestra objetivo con un grupo de edad o grado similares seleccionados de la base de datos, etc., en un tiempo récord (12).

La principal contribución a la LC con este corpus es la aportación de un recurso lingüístico que aún hoy escasea. Este corpus, facilita una muestra de la variedad lingüística infantil, tanto en formato de audio como de texto. Además, dichos textos están enriquecidos con codificaciones, de lo cual se

puede extraer automáticamente todo tipo de información referente a los niveles de sintaxis, semántica, discurso, velocidad, fluidez, errores y omisiones.

El estudio de la adquisición y desarrollo del lenguaje debe basarse en la observación directa de la realidad. Su objeto de estudio es el lenguaje infantil en su manifestación más natural y espontánea, el análisis del lenguaje espontáneo tiene una ventaja sobre las pruebas estructuradas, y es que una conversación con un niño es menos forzada que el típico cuestionario de las pruebas, aunque este también tiene limitaciones, y estas recaen básicamente en el hecho de que los niños no se someten dócilmente a un experimento. El poder inducir a un niño a hablar es algo subjetivo que depende de la personalidad del niño como la del investigador y el grado de empatía que se logre o no, por esto conveniente llevar a cabo más de una sesión, en diferentes espacios y situaciones comunicativas (13).

La longitud media de enunciados (MLU, por sus siglas en inglés) es una de las mediciones del lenguaje que se pueden obtener a través del discurso espontáneo. Su principal objetivo es la obtención de datos sobre los aspectos morfológicos y sintácticos del lenguaje tanto en niños con desarrollo típico como con desordenes de la comunicación (14).

En los últimos años la MLU en palabras se ha considerado muy útil, y su fórmula de cálculo se ha mantenido la misma que la original: número total de palabras dividido por el número total de enunciados producidos. Por otro lado Brown propone la MLU en morfemas, este es menos utilizado débido a que su analisis es más complejo, (15) (16) y a pesar de que se centran en diferentes aspectos lingüísticos, altas correlaciones se han observado entre estas dos medidas. (17) (18)

En relación con la variable MLU en este estudio los valores promedio para MLU palabras fue de 5,22 y MLU morfemas de 5,32, el cual difiere con la investigación realizada en Estados Unidos en el año 2010 cuyo propósito fue estandarizar la MLU en palabras y morfemas en niños de 2 años 0 meses a 9 años, en las edades comprendidas entre 4 años y 4 años 11 meses se determinó un valor promedio de 4,19 para MLU en palabras y 4,66 en morfemas (19).

Esta diferencia radica en la mayor posibilidad de flexión lingüística que el español, lo que ocasiona mayores longitudes ante frases semanticamente semejantes a las del inglés (20).

Por otro lado, un estudio de MLU en español desarrollado en Chile, definió una tabla de proyecciones para el MLU en niños de 18 a 68 meses, según está los niños de edades entre 48 y 59 meses (4 años – 4 años 11 meses) el MLU debe estar entre 3.3 y 4.0 (21). Aunque, en la presente investigación el promedio obtenido es superior a estos valores, dos de las muestras están por debajo de 2,9.

Si el valor de MLU en un niño se encuentra entre 2.0 y 2.9 significa que su desarrollo corresponde a la etapa lingüística "sintáctica" y emite ya estructuras de dos palabras. Se trata de una etapa sintáctica jerárquica inicial, probablemente sin artículos, sin preposiciones o con un número restringido de ellas, aunque sí con sustantivos y verbos (22).

Mientras que si el valor de MLU se encuentra en 3.0 o más, el niño inicia una etapa jerárquica compleja. Aparecen conjunciones, más uso de artículos, de género y número, se usan pronombres personales tiempos verbales y adverbios locativos. Posteriormente, según Acosta, entre los 36 y 42 meses se sigue complejizando la sintaxis con el uso de subordinación (23). La MLU es una medida que aumenta con la edad; sin duda, el niño que empieza a hablar haciendo holofrases tiene una media de longitud de 1.00. Por el contrario, el lenguaje largo y encadenado de los adultos tiene longitudes variables, que generalmente superan el 10 (24).

## CONCLUSIONES

- El trabajo realizado ha dado como resultado un corpus lingüístico del habla infantil espontánea del español, conformado por cerca de 7.500 palabras que cuenta con los archivos de audio y texto electrónicos.

- El corpus cumple con la característica de los corpus modernos de ser formato electrónico, permitiendo el almacenamiento y la manipulación de los datos y su posible intercambio con otros investigadores interesados, pero no cumple con la característica de representatividad, porque aunque su diseño es proporcionado respeto al equilibrio de muestras por género, la diversidad en cuanto a los rangos de edad es baja. Por consiguiente se sugiere ampliar el corpus no sólo en cuanto al número de palabras, sino también incrementando el rango de edad de los participantes e incluyendo una mayor variedad de situaciones comunicativas.

- Este corpus aporta a la LC un recurso lingüístico que aún hoy escasea, además permite la extracción de todo tipo de información referente a los niveles de sintaxis, semántica, discurso, velocidad, fluidez, errores y omisiones.

- La medida de longitud de enunciado determinó que la población estudiada se encuentra en etapa lingüística jerárquica compleja, lo que indica que usan en sus enunciados conjunciones, más artículos, género y número, pronombres personales, tiempos verbales y adverbios locativos.

## Trabajos citados


1. Joan Turruella, Joaquim Llisterri. Diseño de corpus textuales y orales. Filología e informática. 1999.

2. Reyzábal MV. Las competencias comunicativas y linguistivcas, clave para la calidad educativa. Revista Iberoamericana sobre calidad, eficacia y cambio en educación. 2012.

3. Duchan J. Supporting language learning in everyday life: Singular Publishing Group; 1995.

4. Stryker S. The vitalization of symbolic interactionism: American Sociological Association; 1987.

5. Jon F. Miller, Karen Andriacchi, Ann Nockerts. Assesing language production using salt software: A clinician´s guide to language sample analysis Middleton, WI: SALT software LLC; 2015.

6. Turcios RAS. T- Student. Usos y abusos. Scielo. 2015.

7. Salazar MG. Chiede. Corpus de Habla Infantil Espontánea del Español. Linguistica Informatica. 2008.

8. Cuéllar SB. La linguistica de corpus: Perspectivas para la investigación linguistica contemporánea. Forma y Función. 2015.

9. Antonio Briz y Grupo Vales. La transcripción de la lengua hablada. Español actual. 2002.

10. Alvar Esquerra María, Corpas Pastor Gabriel. Criterios de diseño para la creación de corpora. 1994;: p. 31-40.

11. Asier Romero, Irati de Pablo, Aintzane Etxebarria y Ainara. Teorización sobre la construcción de corpus orales en la adquisición del lenguaje. Universidad del pais vasco. 2010.

12. John J. Heilmann, Jon F. Miller, Ann Nockerts. Using Language Sample Databases. Language, speech and hearing services in schools. 2010;: p. 84-95.

13. Dale P. Desarrollo del lenguaje: un enfoque psicolinguistico México: Trillas; 1980.

14. Santos ME, Lynce S, Carvalho S, Cacela M, Mineiro A. Mean length of utterance-words in children with typical language development aged 4 to 5 years. CEFAC. 2015;: p. 1143-1151.

15. Eisenbeiss S. Production methods in language acquisition research. John Benjamins Publishing Company. 2010;: p. 11-34.

16. Eisenberg S, Fersko T UC. The use of MLU for identifying impairment in preschool children: a review. American Journal of Speech, Language Pathology. 2001;: p. 23-42.

17. Arif H, Bol G. Counting MLU in morphemes and MLU in words in a normally developing child and child with language disorder: a comparative study. Dhaka University Journal of Linguistics. 2008;: p. 167-82.

18. Oosthuizen H, Southwood F. Methodological issues in the calculation of mean length of utterance. South African Journal of Communication Disorders. 2009;: p. 76-87.

19. Rice M, Smolik F, Perpich D, Thompson T, Rytting N, Blossom M. Mean Length of utterance levels in 6 month intervals forchildren 3 to 9 years with and without language impairments. Journal Speech Language Hear. 2010.

20. Clemente Estevan RA. Medida del desarrollo morfosintactico. Los problemas de la medición y utilización de la M.L.E. Universidad de Málaga. 1989.

21. Pavez MM. Presentación del índice de desarrollo del lenguaje "Promedio de Longitud de los enunciados". Universidad de Chile. 2002.

22. Herrera M, Pandolfi A. El indice PLE como criterio para analizar el lenguaje infantil. Revista de linguistica teoría y aplicada. 1984.

23. Acosta Rodriguez VM. La evaluación del lenguaje: teoría y práctica del proceso de evaluación de la conducta linguistica infantil: Aljibe; 1996.

24. Briz A. Turno y alternancia de turno en la conversación. Revista Linguistica de Aregentina. 2000;: p. 3-27.

Recibido: PARA USO DE SÍGNOS FONICOS
Revisado: PARA USO DE SÍGNOS FONICOS
Aceptado en: PARA USO DE SÍGNOS FONICOS

Contactar con el Autor:
Natalia Andrea Guerrero Quintero
E-mail: nataguerreroq26@gmail.com