CORPUS LINGÜÍSTICO A PARTIR DE LAS TESIS DE MAESTRÍA EN EDUCACIÓN DE LA UNIVERSIDAD DE NARIÑO PARA EL APOYO HERMENÉUTICO EN LA DETECCIÓN DE SENTIMIENTOS Y CATEGORÍAS EN INVESTIGACIÓN CUALITATIVA.

YENNY ESMERALDA CHIPÚ NARVÁEZ

UNIVERSIDAD DE NARIÑO
FACULTAD DE EDUCACIÓN
MAESTRÍA EN EDUCACIÓN
SAN JUAN DE PASTO
2021

CORPUS LINGÜÍSTICO A PARTIR DE LAS TESIS DE MAESTRÍA EN EDUCACIÓN DE LA UNIVERSIDAD DE NARIÑO PARA EL APOYO HERMENÉUTICO EN LA DETECCIÓN DE SENTIMIENTOS Y CATEGORÍAS EN INVESTIGACIÓN CUALITATIVA.

YENNY ESMERALDA CHIPÚ NARVÁEZ

Proyecto de Tesis presentado como requisito para optar el título de Magister en Educación

Asesor

JESÚS INSUASTI, Ph.D.

UNIVERSIDAD DE NARIÑO
FACULTAD DE EDUCACIÓN
MAESTRÍA EN EDUCACIÓN
SAN JUAN DE PASTO
2021

Nota de Responsabilidad

Las ideas y conclusiones aportadas en este Trabajo de Grado son Responsabilidad de los autores.

Artículo 1 del Acuerdo No. 324 de octubre 11 de 1966, emanado por el Honorable Concejo Directivo de la Universidad de Nariño**.**

Nota de aceptación:

Fecha de sustentación: 23 de noviembre de 2021

Puntaje: 86

DR. ALEJANDRA ZULETA MEDINA

Presidente del jurado

MG. JAVIER MAURICIO LÓPEZ

Jurado

DR. ALEXANDER BARÓN SALAZAR

Jurado

San Juan de Pasto, 23 Noviembre 2021

DEDICATORIA

A Dios por ser la fuerza que cada día me guía, ayudándome a seguir adelante, Permitiéndome culminar con éxito mi tan anhelada Maestría.

A mi hija Yarely, quien es mi mayor inspiración para seguir adelante y luchar por todos mis sueños.

A mi esposo Albeiro, quien fue mi compañero de pregrado, siempre brindándome su apoyo y creyendo en mí para lograr mis metas. Gracias Amor.

A mis padres por estar siempre a mi lado apoyándome en todo momento, quienes con su apoyo y bendición impulsan el alcance de mis propósitos.

A mis hermanos y demás familiares quienes con sus palabras de ánimo me apoyaron para terminar mi maestría.

AGRADECIMIENTOS

A Dios por darme la vida y permitirme uno más de mis grandes anhelos, acompañándome y guiándome en cada paso de mi vida.

Manifiesto mi gratitud especial a la Universidad de Nariño, en especial al programa de Maestría en Educación, por su gran labor informativa y académica.

A la Doctora Gabriela Hernández, Coordinadora del programa de Maestría en Educación, por su gran labor, su valioso conocimiento quien desde el principio ha estado apoyándonos a todos los compañeros con sus consejos y compartiendo su conocimiento.

Al Doctor Jesús Insausti por su apoyo personal y humano en el desarrollo de la plataforma, por compartir su experiencia y amplio conocimiento, por la dedicación y apoyo que ha brindado a este trabajo y por la dirección y el rigor que ha facilitado el proceso evolutivo del proyecto.

A la profesional Andreita Martínez, mi familiar y amiga, por su constante atención y colaboración que me brindó en el desarrollo del proyecto.

Al grupo de investigación Galeras.NET del departamento de Sistemas de la Universidad de Nariño de quienes tuve todo el soporte profesional y logístico para alcanzar el objetivo perseguido y por permitirme acceder a sus recursos computacionales como servidores y software.

A todos mis profesores por sus enseñanzas y compartir sus valiosos conocimientos.

RESUMEN

El presente trabajo está basado en la construcción de un corpus lingüístico a partir de las tesis de la Maestría en educación de la Universidad de Nariño para el apoyo hermenéutico en la detección de sentimientos y categorías en investigación cualitativa. La construcción del corpus se realizará a partir de las tesis desarrolladas para la Maestría de Educación de la Universidad desde el año 2017 hasta la actualidad mediante la selección de una muestra para fines específicos, extraídos de la biblioteca digital de la Universidad de Nariño. El desarrollo inicia con el establecimiento de las bases conceptuales de selección del corpus o de los textos que lo van a componer en el plano conceptual, selección de fuentes, descarga de textos, conversión en formato de texto y la selección rigurosa de la base de alimentación del corpus lingüístico registrados como metadatos en una plantillas para su uso como memoria del corpus lingüístico; una vez se tenga compilado el material base de entrenamiento de análisis lingüístico a partir de un ejercicio hermenéutico de la muestra y creado el corpus se realizará una comprobación de la detección de sentimientos y categorías en los resultados de aplicación de instrumentos de recolección de información para las tesis comprobando la efectividad del corpus desarrollado.

Palabras clave: Hermenéutica, Corpus, tesis, investigación cualitativa, metadatos, análisis de contenido; categorización; investigación; lingüística computacional.

ABSTRACT

The present work is based on the construction of a linguistic corpus from the theses of the Master's in education at the University of Nariño for hermeneutical support in the detection of feelings and categories in qualitative research. The construction of the corpus will be carried out from the theses developed for the Master of Education of the University from 2017 to the present by selecting a sample for specific purposes, extracted from the digital library of the University of Nariño. The development begins with the establishment of the conceptual bases of selection of the corpus or of the texts that are going to compose it in the conceptual plane, selection of sources, downloading of texts, conversion into text format and the rigorous selection of the feeding base of the linguistic corpus registered as metadata in a template for its use as memory of the linguistic corpus; Once the basic material for linguistic analysis training has been compiled from a hermeneutical exercise of the sample and the corpus has been created, a check will be made of the detection of feelings and categories in the results of the application of information collection instruments for the thesis checking the effectiveness of the corpus developed.

Keywords: Hermeneutics, Corpus, thesis, qualitative research, metadata, content analysis; categorization; investigation; computational linguistics.

TABLA DE CONTENIDO

LISTA DE GRAFICAS

LISTA DE IMÁGENES

LISTA DE TABLAS

LISTA DE ANEXOS

INTRODUCCIÓN

El desarrollo científico y tecnológico permite avanzar conjuntamente tanto la informática como el acceso y tratamiento computarizado de textos escritos, transcripciones de diálogos, encuestas abiertas, con una rapidez, fiabilidad y facilidad inimaginable hace algunos años, desde la década de los 60 donde según Caravedo Barrios (1999) se da un florecimiento de los estudios basados en corpus, marcado, en parte, por los cambios paradigmáticos y afectado por la incorporación de los computadores en el ámbito lingüístico (p.35), es decir los corpus informatizados y bases de datos textuales las cuales contribuyen y además facilitan el campo de la lingüística y la investigación cualitativa donde una de las tareas cruciales es el manejo de la información recolectada que permite la construcción de datos; y facilita y agiliza además la tarea de analizar la información sitúa al investigador frente a una gran cantidad de información desordenada, desde donde debe sustraer dentro de toda esa cantidad de datos los que realmente necesita y aportan a su investigación, y, deberá hacerlo de forma metódica y estructurada, para obtener resultados útiles que le lleven a conclusiones válidas y beneficiosas para su investigación. Siguiendo las pautas de la Maestría en Educación de la Universidad de Nariño en relación con la metodología de investigación y producción de conocimiento en el campo educativo como el eje central de una práctica pedagógica transformadora, la construcción de un corpus lingüístico para la detección de sentimientos y categorías en la investigación en el programa, resulta una herramienta de trabajo verdaderamente útil para los alumnos que encaran un proceso de análisis de información cualitativa, para encaminar y realizar procesos hermenéuticos veraces y válidos, permitiendo el análisis de grandes cantidades de datos involucrados. El corpus lingüístico para la detección de sentimientos y categorías es capaz de detectar los conceptos y entidades que aparezcan en el texto y relacionarlos con los sentimientos encontrados, dando una polaridad (o puntuación) a cada entidad y concepto mediante un análisis sintáctico del texto, de forma que fundamenta las reglas para la detección del sentimiento en la estructura de las frases, las relaciones entre los sintagmas y la categoría gramatical de cada palabra procesada son datos muy valiosos y al mismo tiempo, difíciles de extraer, debido, paradójicamente, a la cantidad y diversidad de la información disponible, donde su capacidad de identificar las distintas expresiones de un texto, clasificándolas por polaridad (Sentimientos positivos o negativos) y realizando un análisis de sentimiento lo más fino posible, es definitiva a la hora de permitir

conocer tanto el sentimiento global del texto, como la relación existente entre los distintos conceptos y entidades detectados con las polaridades encontradas en él, conociendo así la valoración del usuario sobre estos conceptos y entidades.

Un corpus lingüístico capaz de detectar conceptos y categorías en un texto y relacionarlos con los sentimientos encontrados arroja datos de gran valor dando una segmentación mediante un análisis sintáctico del texto de forma que, se fundamente en las reglas para la detección del sentimiento, la categorización en la estructura de las frases, las relaciones que se encuentren entre los sintagmas y las categorías gramaticales de cada una de las palabras que se procesan; la capacidad de identificar las diferentes expresiones de un texto y de clasificarlos en sentimientos positivos o negativos a través de un análisis de sentimientos de la manera más exacta posible es definitiva, ya que se requiere clasificar y detectar el sentimiento global de un texto así como la relación existente entre los distintos conceptos detectados por las polaridades encontradas entre frases del texto, para esto, se opta por realizar un corpus lingüístico basado en reglas con el que se pretende obtener un análisis exhaustivo y exacto con un fuerte enfoque lingüístico. Para conseguir este objetivo es necesario construir un corpus y recoger algunos metadatos. De acuerdo a (Meyer, I. & Mackintosh, K, 1996) El corpus se elabora a partir de tesis desarrolladas en la Maestría en Educación hasta la actualidad, ya que “la calidad y representatividad de los textos que van a constituir el corpus especializado resulta un factor de más importancia que la cantidad de textos que han de componerlos” (p.12) es por esto que no se debe realizar una selección arbitraria de fragmentos que se han de incluir ya que se corre el riesgo de omitir alguna descripción conceptual por lo tanto se elegirá un conjunto de textos que proporcione un alcance equilibrado de todos los aspectos temáticos del dominio.

El trabajo se divide en tres partes, la primera se centrará en los objetivos y la metodología aplicada en la elaboración del trabajo, la segunda contemplará la descripción detallada de selección de la base lingüística y el desarrollo y alimentación del corpus lingüístico, y la tercera y última parte se enfocará a la implementación de la aplicación de pruebas de funcionamiento del corpus, tras el diseño y su implementación deberá ser evaluado y comprobar su efectividad, para esto se ejecutara el software a una selección de encuestas para determinar sentimientos positivos o negativos de las respuestas de dichas encuestas. Finalmente se analizarán los resultados buscando problemas de análisis y solucionando los errores hallados a fin de demostrar las posibilidades infinitas que ofrecen los corpus para las investigaciones de tipo cualitativo e

incentivar el uso de estas herramientas e impulsar el desarrollo de estos estudios como una propuesta tendiente al cambio educativo con nuevos enfoques que renueven el quehacer teórico y práctico no solo de la investigación sino de la educación.

CAPITULO I ASPECTOS GENERALES

1.1 Objeto o Tema de Investigación

Corpus Lingüístico en Ciencias de la Educación. Su aprovechamiento en el ejercicio hermenéutico en la investigación cualitativa.

1.1.2 Titulo

Corpus lingüístico a partir de las tesis de Maestría en Educación de la Universidad de Nariño para el apoyo hermenéutico en la detección de sentimientos y categorías en investigación cualitativa.

1.1.3 Área de Investigación

El desarrollo del tema de investigación se enmarcará en el Grupo de Investigación E – TIC en el área TIC en Educación.

1.1.4 Línea de investigación

La presente Investigación se articula en la línea de investigación de Tecnologías de la Información y la Comunicación para la Educación por tratarse de un corpus lingüístico a partir de tesis de la Maestría en Educación de la Universidad de Nariño para la detección de sentimientos y categorías en la investigación cualitativa, el cual se fundamenta en los Ambientes de Aprendizaje mediados por TIC, enfocados a la investigación y el apoyo a las prácticas educativas en diferentes escenarios, teniendo en cuenta que la sociedad del conocimiento, la apropiación y el uso de las TIC cobra mayor importancia gracias a su dinamismo e innovación permanente en los procesos educativos de manera que se fortalezca los procesos investigativos en las Ciencias de la Educación lo cual determina la importancia del desarrollo de un corpus lingüístico para la detección de sentimientos y categorías en las investigaciones de tipo cualitativo y de esta manera incentivar, promover y articular la docencia, la investigación y la extensión en los diferentes niveles educativos estableciendo vínculos con redes de conocimiento, grupos de investigación y comunidades virtuales, entre otros.

1.2 Descripción del Problema

Los estudios cualitativos aportan información sobre las motivaciones profundas de las personas, cuáles son sus pensamientos y sus sentimientos, proporcionan información para adecuar el diseño metodológico de un estudio cuantitativo e información útil para interpretar los datos cuantitativos. De acuerdo a Campoy Aranda & Gomes Araújo (2009) las técnicas cualitativas nos proporcionan una mayor profundidad en la respuesta y así una mayor compresión del fenómeno estudiado, el uso de métodos y técnicas cualitativas ha recibido numerosas críticas basadas en la presunta falta de objetividad, la imposibilidad de reproducción de sus datos, la falta de validez, etc. Estas críticas provocaron en los investigadores cualitativos una posición de inferioridad y de falta de reconocimiento a sus trabajos (Campoy Aranda & Gomes Araújo, 2009), la utilización de unas u otras técnicas dependerá del marco de investigación a realizar, además, intervienen otros factores a considerar como lo son el tiempo disponible, los recursos y la fuente de financiación, el conocimiento previo acumulado sobre el tema específico y el grado de encadenamiento del estudio concreto con otras investigaciones; estos juicios deben ser resueltos con herramientas que permitan un juicio objetivo de la información recolectada, de manera que se utilice eficazmente y proporcione la enorme riqueza informativa que pueden facilitar los diferentes instrumentos de recolección de información para de esta manera controlar y corregir los sesgos propios de cada método; es por esto que el presente trabajo trata la construcción de un corpus lingüístico específico que permita la detección y clasificación de sentimientos que generan las respuestas obtenidas de entrevistas personales, encuestas abiertas, y método de observación de las investigaciones de tipo cualitativo.

La metodología cualitativa según Anguera (1995) se define "como una estrategia de investigación fundamentada en una depurada y rigurosa descripción contextual del evento, conducta o situación que garantice la máxima objetividad en la captación de la realidad, siempre compleja, y preserve la espontánea continuidad temporal que le es inherente, con el fin de que la correspondiente recogida sistemática de datos, categóricos por naturaleza, y con independencia de su orientación preferentemente ideográfica y procesual, posibilite un análisis que dé lugar a la obtención de conocimiento válido con suficiente potencia explicativa." (Anguera, M. T., 1995)

Algunos de los problemas potenciales con los que se enfrentan los investigadores de la maestría de Educación de la Universidad de Nariño que optan por el enfoque cualitativo estriban en que requieren de mucho tiempo para planear, poner en práctica y analizar los datos

recopilados donde se debe además imparcializar la intencionalidad del autor, analizar los contenidos desde el valor del contexto original, y evitar las cuestiones que tienen que ver con la subjetividad en la interpretación, este proceso supone altos costos de tiempo para el investigador; al desarrollar investigaciones hay quienes piensan que están analizando datos, mientras en realidad sólo los están volviendo a describir los datos obtenidos, para un investigador que pasa mucho tiempo en el análisis y clasificación de la información recolectada de datos textos, entrevistas, transcripciones, audio/video, observaciones de campo, así tanto el contexto lingüístico como el resto de elementos externos (factores culturales, y sociales entre otros) que dan forma a una respuesta concreta en una entrevista, encuesta u observación, como los errores gramaticales y de expresión, faltas de ortografía, ausencia o uso incorrecto de puntuación, uso de ciertos lenguajes o "jergas" propias del medio o el grupo de usuarios concreto pueden modificar de forma importante el significado de los elementos de un texto.

Ahora bien, dejando de lado toda la problemática que surge del análisis del lenguaje natural, el análisis de sentimiento por sí sólo también plantea una serie de problemas a resolver como la forma de expresar un sentimiento que generalmente se encuentra "oculta" en el sentido de la frase, por lo que para detectarla se necesitaría un análisis de muy alto nivel, algo muy complejo de lograr. Para realizar análisis sobre textos de encuestas entrevistas u observación, es necesario el uso computacional del lenguaje natural, lo cual implica lidiar con una serie de dificultades debidas a la propia naturaleza del lenguaje natural humano ya que la interpretación del sentimiento de un texto no es sencilla dado que el lenguaje natural presenta muchas características que lo hacen un verdadero reto para las ciencias de la computación, aspectos como la ambigüedad, la espontaneidad, la falta de fluidez, o las referencias y abreviaturas, quedan muy lejos de ser completamente definidas por las disciplinadas reglas que rigen el comportamiento de un sistema informático; sin embargo el corpus lingüístico, es capaz de desarrollar sistemas que pueden trabajar con lenguaje humano de una forma efectiva, ya sea recibiendo instrucciones en lenguaje natural, produciendo una interfaz de respuesta conversacional o procesando y realizando tareas de análisis de lenguaje humano, como es el caso que nos concierne para nuestro sistema.

De aquí, la necesidad de constituir un corpus lingüístico funcional para el análisis de datos de una investigación, que permita que los datos obtenidos a través de observaciones, encuestas y entrevistas se facilite mediante el corpus organizando los resultados e interpretaciones de sentimientos positivos o negativos identificando rápidamente patrones y tendencias, posibilitando

no solo la exploración de datos de forma rápida y eficaz sino, el comparar y decodificar los análisis por categorías de forma confiable y en corto tiempo ya que, un corpus lingüístico permite a cualquier investigador generar uniformemente datos de espectros de masas de la máxima calidad, de forma sistemática, sin necesidad de ninguna formación ni conocimientos especiales, comprender sentimientos positivos y negativos y actitudes de los encuestados que no son cuantificables ya que este tipo de datos tienen como principal característica que no se pueden medir, ni expresarse con número, deben ser interpretados, por lo que su uso es muy importante para fundamentar cualquier investigación analizando datos cualitativos para responder incógnitas sobre el "cómo" y el "porqué" de una situación, en vez de "cuántos". En resumen, es de gran ayuda para economizar tiempo (sobre todo con conjuntos de datos extensos), pero también le permite dedicarse a profundidad en sus datos de investigación, que, por cuestiones de tiempo, a veces se desaprovechan sirviendo de soporte al investigador y fortaleciéndolo durante su trayecto de investigación y de análisis.

En lo que se refiere al interés personal, el desarrollo del presente trabajo permite poner en práctica lo aprendido en la carrera profesional y además enfocado y aportando de manera significativa con una herramienta a los investigadores de la Maestría en Educación de la Universidad de Nariño.

1.2.1 Delimitación del Problema

1.2.1.1 Delimitación Conceptual.

La parte conceptual estará delimitada por el estudio que se haga del concepto de Corpus lingüístico en Ciencias de la Educación; su aprovechamiento en el ejercicio hermenéutico en la investigación cualitativa.

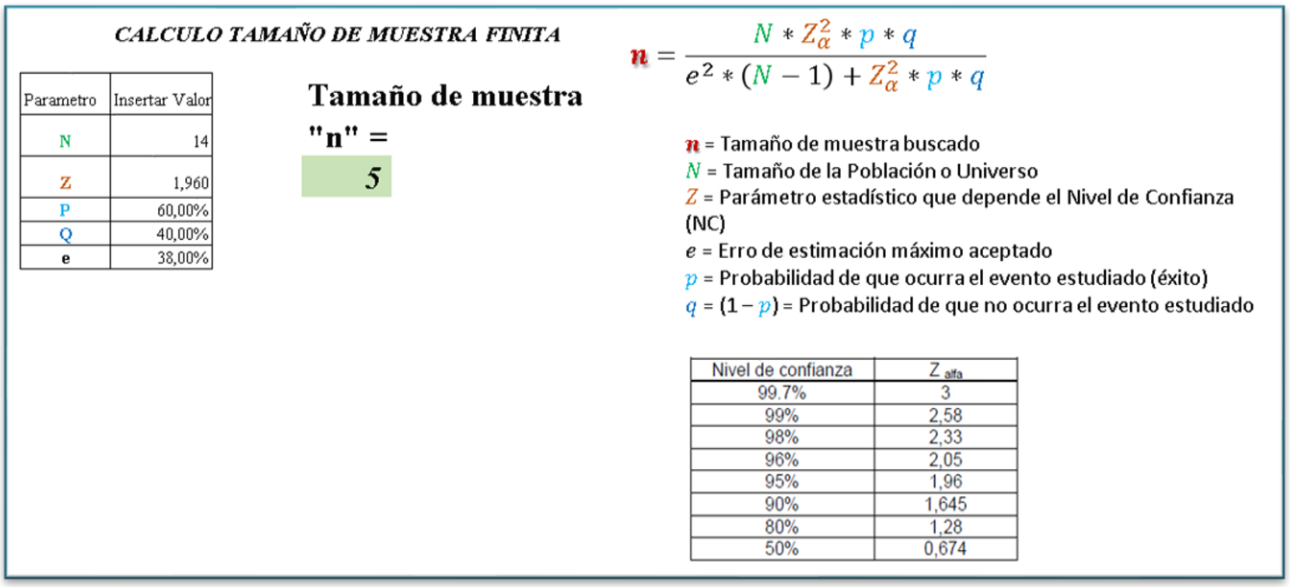
1.2.1.2 Delimitación Espacial.

La delimitación se enmarca en la Maestría en Educación de la Universidad de Nariño donde se observará el impacto del Corpus lingüístico en los resultados de la aplicación de instrumentos de recolección de información de las tesis desarrolladas en la Maestría.

1.2.1.3 Delimitación Temporal.

El proceso de investigación de los trabajos de grado elegidos para el desarrollo del Corpus lingüístico comprenderá un periodo de tiempo que va desde el año 2017 hasta el año 2021. Esta selección temporal se basa en el cálculo de muestra representativa desde el inicio de la maestría (2008) hasta el presente con un nivel de confianza de 95% y un margen de error de 40% mediante la fórmula:

Imagen 1 Cálculo de muestra



Fuente: Esta Investigación.

1.2.2 Formulación del Problema

¿Cómo apoyar tecnológicamente los procesos hermenéuticos en la detección de sentimientos y categorías en resultados de aplicación de instrumentos de recolección de información en investigación cualitativa?

1.3 Objetivos de la Investigación.

1.3.1 Objetivo General

Establecer un Corpus lingüístico a partir de las tesis de Maestría en Educación de la Universidad de Nariño para el apoyo hermenéutico en la detección de sentimientos y categorías en investigación cualitativa.

1. 3.2 Objetivos Específicos

1. Establecer bases conceptuales de selección del corpus o de los textos que lo van a componer en el plano conceptual

2. Desarrollar una codificación y programación del material de entrenamiento de análisis lingüístico a partir del ejercicio hermenéutico de una muestra de las tesis de Maestría en Educación hasta la actualidad.

3. Realizar el procesamiento informático de los textos para la materialización e implementación del soporte lógico y físico (hardware y software) del corpus.

4. Aplicar pruebas de validación de funcionamiento de corpus en la detección de sentimientos y categorías en los resultados de aplicación de instrumentos de recolección de información para las Tesis de la Maestría.

1.4. Justificación

En la literatura especializada se abordan dimensiones sustanciales del proceso de indagación científica, pero no siempre se realizan apuntes o se sugieren pistas para enfrentar los dilemas que supone el manejo de grandes cantidades de información y su análisis imparcial, a la luz de las competencias teórico-metodológicas que esto demanda es importante que los docentes se apropien de las tecnologías para el diseño de ambientes de aprendizaje colaborativo que faciliten el acompañamiento pedagógico y el desarrollo de investigaciones de los estudiantes que inician investigaciones de tipo cualitativo entender la naturaleza y posibilidades del uso de un corpus que simplifique el análisis de sus investigaciones cualitativas para favorecer los procesos de

construcción de conocimiento; por ello, debe apelarse a una didáctica de la investigación cualitativa donde los procesos de transmisión-apropiación se centran en quehaceres y operaciones de la actividad científica que se apoyen en las tecnologías de la información y la comunicación (TIC) que han pasado a ser recursos que favorecen la configuración de espacios del quehacer investigativo ya que es un recurso educativo abierto para uso y beneficio de la comunidad educativa mundial; particularmente para su utilización por parte de profesores, alumnos e investigadores en diversos niveles educativos.

En términos generales, un corpus lingüístico se puede entender como la recopilación de textos de fuentes reales con la finalidad de detectar sentimientos y categorías en resultados de aplicación de instrumentos de recolección de información. Hoy en día los corpus están en formato electrónico para poder procesarlos con algún programa de ordenador. Una definición más completa es la que nos ofrece Victoria López Sanjuán (2017), tutora de la UNED para el Departamento de Filologías Extranjeras y sus Lingüísticas: “Un corpus es un conjunto de textos recopilados con una finalidad lingüística, que sirve como fuente de información para demostrar algún aspecto concreto de la lengua a la que pretende representar y que facilita que se hagan generalizaciones a partir de los datos que hay en su contenido”. (López Sanjúan , 2017)

Etimológicamente el corpus es un conjunto de textos de materiales escritos o hablados, agrupados bajo un conjunto de criterios mínimos para realizar análisis lingüísticos ya sean de tipo cualitativo o cuantitativo, para lo cual su base debe tener una representatividad, variedad y equilibrio adecuados para el fin que se busca; esencialmente el corpus lingüístico que se pretende desarrollar es una recopilación de textos de fuentes reales de trabajos de grado y disertaciones desarrolladas en la Maestría en Educación hasta la actualidad con la finalidad de detectar sentimientos y categorías; lo que permitirá construir un corpus veraz, no datos inventados o ajustados a los juicios personales de los investigadores, además facilitaría el procesamiento de la información recopilada de forma imparcial y eficaz en comparación con los estudios manuales que se vienen realizando hoy por hoy, esto permite la obtención de afirmaciones más objetivas que las que se consiguen con la introspección y juicio propio del investigador dado que el programa “entiende e interpreta” la exactitud del contenido.

Son diversas y de gran valor las propuestas de análisis de las que se puede hacer acopio de aspectos que ayudan al docente o al estudiante - investigador que pretende interpretar y analizar los datos de una investigación; sin embargo, pocas veces se queda satisfecho de los resultados en

la aplicación de estos estudios, pues, con frecuencia representan un punto de vista parcial de la totalidad significativa que se puede encontrar; actualmente, en muchas ramas de la educación y sobre todo en investigación de tipo cualitativo, procura trabajar con datos reales y lo más exhaustivos posibles que permitan reproducir con la máxima veracidad las características del objeto de estudio, esto implica que, de algún modo, hay que recopilar, en cantidades más o menos grandes, muestras de los elementos que constituyen la realidad que se quiere investigar, lo cual provoca que el investigador se pueda encontrar delante de cantidades considerables de documentos que aportan un número de datos tan grande que sólo una codificación, ordenación y organización de estos datos en la proporción adecuada pueden probar la detección de sentimientos y categorías en los resultados. La búsqueda del significado es también "el objetivo de toda hermenéutica" (Beuchot, 2008) y esta es "la disciplina de la interpretación de textos" (Beuchot, 2008). La hermenéutica está en una permanente búsqueda de equilibrio en las interpretaciones, como se observa en la hermenéutica analógica que se centra en la proporción, entre la pretensión de claridad (la univocidad) y la oscuridad y confusión en la interpretación (equivocidad) (Beuchot, 2008). No obstante, aunque la hermenéutica analógica considera los niveles de análisis lingüístico sintáctico, semántico y pragmático, así como el contexto del texto, y alude a la interpretación de niveles micro y macro, es relevante hacer notar que el punto de partida está en el lector del texto (Grondin, 2018) distingue al menos cinco "vías" para llegar al sentido que anima la hermenéutica: la sensibilidad, el significado, la dirección, la inteligencia y lo razonable. De todas ellas, aparentemente, la sensibilidad es la más valiosa porque es el primer significado de "sentido" de los textos. Como dice Grondin (2008):

El primer significado de sentido designa o caracteriza una capacidad de sentir las cosas, una sensación de algo. El mejor ejemplo para ilustrar esto es cuando se habla de los cinco sentidos: tocar, oler, mirar, gustar, escuchar. Somos seres abiertos al mundo -esta es una idea clásica de la filosofía- y estamos abiertos a él gracias a nuestros cinco sentidos. Entonces, el sentido hermenéutico, como quiere entenderlo, primeramente, es un sentido sensitivo: una capacidad de sentir las cosas. De esta manera, la hermenéutica es una especie de sensibilidad, una sensibilidad para algo. (Grondin, 2018, págs. 17-33)

Un corpus lingüístico facilita enormemente la tarea mecánica de editar y organizar el texto en formato electrónico, y también permite un acceso rápido y selectivo a la información, condición que no cumplimos cuando buscamos manualmente toda la información obtenida con papel y lápiz

y son cualitativamente Además, el proceso de organización del corpus lingüístico es más flexible de forma sistemática y clasificada, ya que toda la información, documentos originales, citas, descripciones de códigos y comentarios se almacenarán en el programa y se podrá acceder de forma inmediata, que además reduce la cantidad de investigación la cantidad de papel consumido por el personal y el tiempo que lleva buscar y clasificar los datos, de ahí la importancia de obtener un corpus suficientemente organizado y representativo para que pueda ser explotado con ciertas garantías de éxito que permita analizar los sentimientos que antes solo eran conjeturas o especulaciones provenientes de impresiones más o menos fundadas de la lingüística. Además, el desarrollo de este proyecto de investigación además servirá como medio de consulta para futuras investigaciones relacionadas con la temática.

CAPÍTULO II. MARCO REFERENCIAL

2.1 Antecedentes de la Investigación

Indagación bibliográfica en investigaciones anteriores, tanto en el ámbito nacional, regional como en el internacional.

2.1.1 Internacionales

Título: Corpus lingüístico y tecnologías para la enseñanza, el aprendizaje y la investigación en traducción audiovisual y accesibilidad lingüística (subtitulado para sordos, audio descripción para ciegos y lengua de signos española): normativas de aplicación.
Autores: Rica Peromingo, Juan Pedro, Rodríguez Redondo, Ana Laura, González Sánchez, Mª Carmen, Fernández Lijó, Gloria, López, Mª Monteagud, Puchol Vázquez, Blanca, Orero Clavero, Pilar, Matamala, Ana, Díaz Cintas, Jorge, Sáenz Herrero, Ángela De Higes Andino, Irene, Andrades Moreno, Arsenio, Martínez Portillo, Sara, Bolaños García-Escribano, Alejandro, Mata Pastor, Manuel, Soroa Sainz, Paloma
Universidad: Universidad Complutense de Madrid (UCM)
Año: 2019
Objetivo General: ✓ Consolidar y mantener las actividades encaminadas hacia la innovación educativa y la construcción del EEES para la enseñanza de la TAV y de la lengua y lingüística por parte de un grupo innovador de profesores, que trabajan juntos desde hace varios años y la incorporación de nuevos investigadores. ✓ Incorporar profesores innovadores pertenecientes a distintos centros y universidades, tanto nacionales como internacionales, con el fin de lograr mayor difusión de nuestros resultados. ✓ Potenciar los recursos de las universidades implicadas generando un portal educativo, disponible para todo el personal académico, que sea de acceso abierto y gratuito a través de Internet, y hacerlo accesible a otros contextos educativos universitarios.
Objetivos específicos:

- ✓ El objetivo principal es el de continuar con el CALING en la confección de materiales de innovación docente y de evaluación para la enseñanza de la TAV en los programas universitarios de filologías y de TeI, en concreto en lo que respecta a la accesibilidad lingüística. Se tomará como base la lengua inglesa. Se constituirán subgrupos de trabajo y se intentará poner en común los trabajos individuales para el diseño y desarrollo de los materiales y los recursos específicos en la enseñanza de las modalidades en TAV anteriormente mencionados. Puesto que el proyecto está concebido como una continuación de los PIMCD anteriores (PIMCD 59, PIMCD 30, PIMCD 9 y Proyectos Innova 4 y 6), se podrán poner en práctica en otros contextos educativos de nueva incorporación (UCL, UJI) las actividades ya confeccionadas durante los cursos académicos anteriores y que figuran en la página web del equipo de investigación (http://avlearningarchive.com/). Se elaborarán nuevas actividades con el fin de poder establecer unas pautas de enseñanza y aprendizaje de las cuestiones lingüísticas sobre accesibilidad más relevantes en el mundo de la TAV y sugerir modificaciones en la normativa actual sobre SPS y AD (Normas UNE 153010 y 153020). Las actividades confeccionadas en años anteriores y las nuevas se seguirán testando con los especialistas en SPS, lengua de signos española (LSE) y AD y con los receptores sordos (a través del CSIM y de la Oficina para la Integración de Personas con Discapacidad (OIPD) de la UCM) y ciegos (a través de la UAB y la ONCE).
- ✓ Un segundo objetivo del proyecto es entablar relaciones académicas con centros especializados en la traducción audiovisual y la mediación lingüística para hacer disponibles a sus usuarios los materiales, herramientas, tecnologías y evaluación de materiales que el equipo de investigación recopile: la ONCE (www.once.es/new), el CESyA (Centro Español de Subtitulado y Audio descripción: www.cesya.es/), el NLSE (www.cnlse.es/) o el departamento de SPS y AD de RTVE. Asimismo, se mantendrá una línea abierta con ATRAE, única asociación española sobre TAV, para contrastar las realidades del sector y estudiar la posibilidad de publicaciones conjuntas y se colaborará con MINCASOR Deaf Film Company (www.mincasor.com/), asociación que organiza anualmente la Muestra Internacional de Cine de Sordos.
- ✓ El objetivo final de todo este proyecto es seguir compilando materiales y evaluaciones para el CALING —que se añaden a los ya existentes— con el uso de las

nuevas TIC que puedan ser aplicables en el futuro no sólo a los estudios de TeI de la UCM, de la UAB, del UCL, de la UNED y de la UJI, sino también a los estudios de lingüística (inglesa, en un principio). Está previsto hacer públicos los resultados a través de publicaciones, presentaciones en congresos nacionales e internacionales, publicación de las herramientas, metodologías y tecnologías tanto en el Campus Virtual de la UCM como en aquellas plataformas virtuales de las universidades interesadas en este estudio metodológico, con el fin de fomentar el aprendizaje autónomo de los estudiantes en la traducción de materiales audiovisuales.

Conclusiones:

- ✓ El proyecto continúa con la recopilación del corpus lingüístico (CALING - Corpus de Accesibilidad Lingüística) de materiales y evaluaciones de receptores reales (sordos y ciegos) comenzado en el proyecto de innovación anterior para la enseñanza-aprendizaje de la accesibilidad lingüística en la TAV en un contexto universitario. Como se puede observar con los resultados presentados en este trabajo, se hace necesaria una enseñanza más directa y explícita de la accesibilidad lingüística en las clases de traducción audiovisual. En el contexto del corpus CALING los estudiantes que participan son universitarios con un nivel avanzado de la lengua inglesa (C1, según el marco europeo de las lenguas).
- ✓ Los estudiantes han realizado actividades de subtitulado para oyentes en las clases con anterioridad, pero no han recibido hasta el momento de la práctica en las clases una enseñanza explícita en las cuestiones técnicas específicas para el subtitulado para sordos y el audio descripción para ciegos ni en las cuestiones lingüísticas propias de estas dos modalidades. Los estudiantes han recibido en las clases una enseñanza directa de las dos Normas UNE que regulan estas modalidades de TAV y han seguido escrupulosamente las indicaciones que ahí se reflejan. Los receptores sordos y ciegos que han participado en las evaluaciones tienen conocimiento de (y, en rasgos generales, apoyan) las dos Normas UNE de trabajo, por lo que sus comentarios de la producción de los estudiantes siempre han tenido en cuenta el conocimiento que tienen de las reglas que rigen el SPS y la AD en la televisión y el cine.
- ✓ Se pueden extraer varias reflexiones de los resultados preliminares presentados en

la sección 5 de este artículo. Con respecto al SPS, los receptores han concluido que el ritmo de lectura de los subtítulos debe ser más lento, prefieren en general el uso de más colores que los 4 básicos para los 4 personajes principales más el color blanco para el resto de personajes, están conformes con el posicionamiento de los subtítulos y la información contextual en la parte inferior de la pantalla centrados, agradecen la literalidad de los subtítulos y no están muy conformes con el tipo de información contextual que se ha reflejado en lo que respecta los sentimientos, emociones y estados de ánimo de los personajes, puesto que no les queda claro y creen que existe un uso excesivo de este tipo de información. En general, los resultados indican que los informantes aceptan en gran medida las indicaciones que se incluyen en la Norma UNE 153010, pero sugieren unos cambios importantes que se deberían contemplar en futuras revisiones de la normativa, especialmente en lo que respecta a qué tipo de datos se incluye en la información contextual en el subtitulado para sordos, cómo se presentan esos datos y cuántos colores se deberían utilizar, aparte de los 4 principales (más el blanco para el resto de los personajes).

- ✓ Con respecto al otro tipo de actividades analizadas por receptores reales, las AD, nos encontramos con un caso similar al anterior: los receptores están en general conformes con la producción de los estudiantes, aunque paralelamente muestran su disconformidad con algunos aspectos que han criticado: en primer lugar, consideran que el texto audio descrito es excesivo en muchas ocasiones, lo que provoca gran estrés auditivo y mental; en segundo lugar, han valorado positivamente (siguiendo las indicaciones de la Norma UNE 153020) el hecho de que la voz de los estudiantes audio descriptores suene natural, hayan vocalizado correctamente, no hayan utilizado cacofonías en el texto y hayan utilizado una entonación apropiada y no monótona. Echan en falta, de todos modos, que haya una mayor «implicación emocional» por parte de los audio descriptores, aspecto este que iría en contra de las recomendaciones de la Norma UNE de trabajo. Relacionado con este aspecto es también llamativo que los evaluadores hayan «pedido» que la neutralidad que se pide a los audio descriptores desaparezca para que los textos audio descritos ganen en naturalidad. También han insistido en que, en general y no solo con respecto a las AD que han evaluado de los estudiantes sino también las AD profesionales que consumen en televisión o en el cine, las AD deben realizarse a una velocidad más lenta para no perder tanto texto audio descrito puesto que no les da tiempo a procesar toda la información que

se incluye en la descripción.

✓ Con estos resultados preliminares se hace necesaria una investigación más completa y amplia sobre cómo afrontar la enseñanza y aprendizaje de la accesibilidad lingüística en el aula (Assis Rosa 2016, De Higes y Cerezo 2018, Matamala y Orero 2016, Neves 2016, Remael et al. 2016, Rica Peromingo 2016) que incluya receptores reales que puedan evaluar la recepción de SPS y AD que se emiten en televisión, cine o DVD o las que realizan los propios estudiantes, como es el caso del corpus CALING.

✓ Con estudios como el que aquí se presenta se puede incidir en el aula en aquellos aspectos que los evaluadores han considerado primordiales. Al mismo tiempo, consideramos fundamental continuar la evaluación de las Normas UNE 153010 y 153020 para una mejora que se adapte a las necesidades de los receptores sordos y ciegos que consumen SPS y AD en un contexto español. Este aspecto nos anima a continuar con la recopilación de materiales y evaluaciones audiovisuales para la enseñanza y el aprendizaje de las modalidades de TAV identificadas con la accesibilidad lingüística dentro del marco del proyecto de investigación y del corpus CALING. Nuestra intención más inmediata, como se indicó en la introducción de este artículo, es recopilar durante los siguientes cursos académicos más actividades y ampliar el número de receptores reales que nutran nuestro corpus de datos significativos y amplios que nos permitan llegar a generalizaciones más fiables y completas. Esto repercutiría, por un lado, en una mejora en la enseñanza de estas dos modalidades de TAV en el aula y, en segundo lugar, en una mejor calidad de los SPS y las AD que se emiten en televisión y en el cine.

✓ Finalmente, se pretende una recepción óptima por parte de los colectivos sordos (o con discapacidad auditiva) y ciegos (o con discapacidad visual) de materiales audiovisuales en las mismas condiciones que los oyentes y videntes. Consideramos, finalmente, que es necesaria más investigación en este campo que tenga en cuenta la recepción de estas modalidades de TAV por receptores reales puesto que aunaríamos los conocimientos más teóricos con las necesidades reales de aquella población a la que van destinados los productos subtitulados para sordos y audios descritos para ciegos. (2019)

Aporte al trabajo de investigación que se está realizando

El aporte que realiza este trabajo es la elaboración de nuevos materiales, recursos específicos y procesos de evaluación para la puesta en práctica de corpus a las nuevas

tecnologías y metodologías de enseñanza y aprendizaje de la TAV, con la recopilación del corpus CALING (Corpus de Accesibilidad Lingüística).

Título: Los corpus lingüísticos al servicio de la semántica: su empleo en la delimitación de sentidos contextuales.
Autores: Martín Padilla, Kenia
Universidad: Universidad Autónoma de Barcelona
Año: 2015
Objetivo General: Presentar un modelo de análisis léxico en familias de palabras basado en la existencia de una matriz semántica común o significado invariante, que se mantiene constante tanto en la variación gramatical de la raíz, como en su variación denotativa. Tomando como ejemplo la familia de palabras español, el propósito es comprobar cómo la información proporcionada por los corpus generales posibilita la delimitación de los distintos sentidos contextuales que las unidades adquieren en el uso.
Objetivos Específicos ✓ Ofrecer una propuesta de sistematización del léxico. ✓ Establecer una clara distinción entre la significación idiomática y la significación denotativa, articulando un modelo que logre describir ambas.
Conclusiones: ✓ Al analizar los términos de manera conjunta, además de facilitar la obtención de resultados generales aplicables a un gran número de formas, el método de análisis en familias de palabras arroja un rayo de luz sobre uno de los más oscuros escollos del estudio del vocabulario: permite estudiar el léxico como un conjunto cerrado de KENIA MARTÍN PADILLA Scriptum Digital Vol. 4 (2015), pp. 165-185 180 unidades. Así, como la estructura organizativa queda establecida por la propia lengua, los esfuerzos han de centrarse en observar qué rasgos caracterizan la raíz. El concepto de significado empleado, por otra parte, asume la naturaleza plural del aspecto semántico de las unidades léxicas al entenderlas como entidades compuestas, que son el resultado de la articulación de información lingüística jerarquizada en distintos niveles. Esta manera de observar la realidad significativa como una red en la que se superponen elementos de distinta

naturaleza permite describir no sólo la significación primaria o léxica, sino considerar también la significación gramatical, y no sólo la dimensión lingüística o idiomática, sino también la significación denotativa y designativa.

- ✓ Separar estas dos dimensiones ayuda a determinar qué rasgos aporta el contenido léxico y qué rasgos proceden del contorno. Para acometer la labor de estudiar los aspectos contextuales el empleo de corpus se convierte en una excelente herramienta, al permitir analizar la variación denotativa de acuerdo con sus propiedades combinatorias y designativas, atendiendo al uso real de las formas. Asimismo, la existencia de corpus diacrónicos supone un auténtico adelanto en el estudio histórico de las unidades porque posibilita la datación aproximada de los sentidos y permite desentrañar aspectos culturales. En un estudio de estas características, estas cuestiones resultan de gran interés pues, como se ha comprobado, los sentidos desaparecidos pueden explican usos actuales. La consideración de todos los factores que han podido intervenir en la fijación de sentidos y su rastreo minucioso en las fuentes ayuda enormemente a la hora de clasificar y ordenar el material.
- ✓ Además de facilitar la evidencia de uso, los corpus empleados aportan a este estudio dos ventajas primordiales: proporcionan un conjunto amplio de usos concretos de una unidad particular, y permiten el acceso inmediato a una amplia selección de textos de diferentes épocas, tipologías y áreas temáticas. Sin embargo, presentan también algunas limitaciones. La más evidente es que constituyen una selección, mayoritariamente de textos escritos y de registros cultos, lo que implica que muchos usos propios de registros coloquiales, del lenguaje técnico o especializado, o de determinadas diatopías, pueden no recogerse. A esto habría que unir el hecho de que, en un estudio como el que proponemos, la ausencia de etiquetado semántico no permite acceder de forma automatizada a los distintos sentidos que posee una misma unidad. El etiquetado semántico, que se ha llevado a cabo en otros corpus (p.e. ADESSE) supone el gran reto para el futuro.
- ✓ El inconveniente es la complejidad de las relaciones semántico-denotativas que presentan las unidades, que dificulta a la lingüística computacional resolver los problemas relacionados con la creación de nuevos sentidos formados sobre la base de metáforas y metonimias. Sin embargo, para llevar a cabo la informatización de los datos, sería de gran ayuda contar con estudios semánticos previos. En este sentido, la retroalimentación parece

una opción plausible: los estudios lingüísticos basados en corpus pueden alimentar la creación de nuevos corpus anotados. (2015)
Aporte al trabajo de investigación que se está realizando Este trabajo realiza un aporte de la perspectiva potencialmente fructífera de las posibilidades que ofrecen los corpus lingüísticos, la extracción de información semántica frente a las dificultades del propio análisis semántico al presentar un modelo de análisis léxico en familias de palabras basado en la existencia de una matriz semántica común o significado invariante, que se mantiene constante tanto en la variación gramatical de la raíz, como en su variación denotativa comprobando cómo la información proporcionada por los corpus generales posibilita la delimitación de los distintos sentidos contextuales que las unidades adquieren en el uso.

Título: ¿Cómo hablan los jóvenes? Los corpus lingüísticos como base para la reflexión teórica y el aprendizaje de lenguas
Autores: Borreguero Zuloaga, Margarita, Hidalgo Downing, Raquel Ángela, Gil Valdés, María Jesús, Martín Gascueña, Rosa, Sancho Pascual, María, Guijarro Sanz, María Nappi, Paolino.
Universidad: Universidad Complutense de Madrid
Año: 2021
Objetivo General: Utilizar los corpus lingüísticos como herramientas para el aprendizaje lingüístico y la reflexión metalingüística.
Objetivos específicos ✓ Reflexionar sobre la variación lingüística a través de los corpus. ✓ Valorar los corpus lingüísticos como herramientas para el aprendizaje de una lengua extranjera. ✓ Reflexionar sobre la importancia de los corpus como base empírica en la investigación en lingüística. ✓ Estudiar las características del lenguaje juvenil en sus distintos niveles de articulación lingüística. ✓ Proponer a los alumnos la recogida de un pequeño corpus para hacerles entender los problemas metodológicos y teóricos que dicha actividad plantea.

Conclusiones:

✓ Una selección de los trabajos realizados en el marco de este proyecto de innovación docente se presentó en una jornada celebrada en modalidad híbrida (presencial y online) en el Salón de Grados de la Facultad de Filología el 10 de marzo de 2021.

✓ Por la situación de emergencia sanitaria esta jornada no pudo celebrarse en abril de 2020, tal y como estaba prevista, y tuvo que posponerse durante un año. En el encuentro participaron los profesores que habían tutorado las investigaciones llevadas a cabo en el marco del proyecto. La mayoría de los estudiantes acudieron presencialmente a presentar sus análisis, aunque también hubo otros que prefirieron presentar a través de la plataforma Zoom o que mandaron un vídeo con sus presentaciones pre-grabadas.

✓ Los asistentes (unos 40 aproximadamente) se conectaron a través de la plataforma Zoom. En todos los casos, se trató de análisis interesantes y rigurosos, con una sólida base lingüística y sociológica en la mayoría de los casos. Merece la pena destacar que la mayoría de los alumnos llevó a cabo sus trabajos cuando eran alumnos de 1º curso de Grado. Para ellos ha constituido un primer acercamiento a la metodología de estudio lingüístico a partir de corpus de gran interés para ellos ya que reflejan su propia variedad lingüística: la lengua juvenil. (2021)

Aporte al trabajo de investigación que se está realizando

Existen en la actualidad numerosos corpus monolingües y multilingües, de lengua hablada y escrita, corpus de textos digitales, de corte sociolingüístico, de hablantes nativos y de aprendices, el aporte de este trabajo al a investigación se centra la utilización de los corpus lingüísticos como herramientas para el aprendizaje lingüístico y la reflexión metalingüística que permite reflexionar sobre la variación lingüística a través de los corpus como base empírica en la investigación en lingüística y entender los problemas metodológicos y teóricos que un corpus lingüístico plantea.

Título: Más allá del Corpus: Big Data en la Investigación Lingüística. Evolución, Análisis y Predicción del uso de la Lengua a través de Twitter.

Autores: Adela González Fernández

Universidad: Universidad De Córdoba España
Año: 2016
Objetivo General: Demostrar la veracidad de la hipótesis y comprobar que no solo es posible utilizar *Big Data* para investigar el lenguaje y las lenguas, sino que, además, esta metodología supone una mejora con respecto al trabajo lingüístico tradicional porque aporta más información que hasta ahora, con una menor inversión de tiempo y de esfuerzo por parte del investigador.
Objetivos específicos ✓ Crear una herramienta que permita la obtención de la información textual proveniente de Twitter, así como su almacenamiento y gestión, para un posterior análisis que tendrá que ser complementado, inevitablemente, por el lingüista. ✓ Desarrollar varios estudios que sirvan como muestra de las numerosas aplicaciones que se pueden derivar del trabajo con la herramienta y la metodología que presentamos y que demuestren la utilidad de *Big Data* en la investigación lingüística. ✓ Demostrar la conveniencia de la unión entre la Lingüística y la Informática y de la apuesta por la innovación en la investigación lingüística.
Conclusiones: ✓ En este ejemplo de investigación, hemos podido demostrar la eficacia de *Wordics Archive* a la hora de llevar a cabo estudios de lingüística aplicada útiles no solo para obtener información relacionada con el lenguaje y su uso, sino también para poder enfocarla a otros aspectos prácticos del ámbito profesional, como el sector de la traducción. Aunque la herramienta devolvía todos los tuits publicados en cada una de las *bounding boxes* descritas, hemos considerado oportuno mencionar los cuatro idiomas más utilizados por los usuarios de Twitter para comunicarse y englobar así el resto en una sola categoría. ✓ Desde un punto de vista profesional traductológico comprobamos, por ejemplo, que el inglés predomina como segunda lengua más usada en todas las capitales en las que no actúa como idioma oficial. Londres, lógicamente, es la excepción. Además, es llamativa la diferencia en esta ciudad entre el uso del inglés y el del resto de los idiomas. Parece claro que el nivel de plurilingüismo en una capital cuyo idioma oficial es el inglés es mucho menor que en el resto de las ciudades, en las que esta lengua figura siempre en segundo

lugar.

- ✓ Los datos dejan pocas dudas acerca del carácter vehicular y universal de este idioma. Por otro el lado, es curioso también el hecho de que, en las capitales estudiadas, el alemán solo se habla de forma significativa en Berlín, mientras que otros idiomas, como el español, el árabe, el francés o el portugués, tienen más presencia en los países europeos. De hecho, estos cuatro, junto con el inglés y el alemán son los únicos seis idiomas registrados como los más utilizados en estas cinco ciudades. Mención especial merece el uso del neerlandés, ya que se trata de uno de los idiomas oficiales de Bruselas. Otros idiomas, sin embargo, que podrían presentar oportunidades para el mundo de la traducción, ya sea por su proximidad con los países estudiados o por el gran número de hablantes que tienen no dejan rastros relevantes en estas ciudades.
- ✓ Observamos que, mediante la aplicación de *Big Data* al campo de la Lingüística, podemos obtener una visión global y realista de las necesidades traductológicas de un lugar concreto en una fecha determinada. Huelga decir que a ningún profesional del sector se le escapa cuáles son los idiomas más utilizados en las distintas capitales. Sin embargo, puesto que los idiomas son un ente vivo y se encuentran en constante evolución, su situación es susceptible de cambio conforme pasa el tiempo o cambian los factores sociales, políticos o económicos. Una investigación de este tipo no solo ahorra tiempo y costes con respecto a los métodos tradicionales, sino que nos brinda la posibilidad de obtener un fotograma del estado de cualquier idioma en el momento que deseemos e incluso, también, en tiempo real. (2016)

Aporte al trabajo de investigación que se está realizando

Gracias a investigaciones de este tipo, es posible establecer las tendencias predominantes del Big Data al campo de la Lingüística en los procesos de formación de palabras nuevas y conocer sus mecanismos de formación, además de contribuir a la actualización de las teorías actuales sobre la formación de neologismos y aportar evidencias para la actualización de obras lexicográficas necesarios en el momento de determinar y clasificar sentimientos y categorías de corpus lingüísticos para fines específicos.

Título: Análisis de contenido y lingüística computacional: su rapidez, confiabilidad y perspectivas
Autores: Brenda Lía Chávez, Jorge Martín Yamamoto
Universidad: Pontificia Universidad Católica del Perú
Año: 2019
Objetivo General: Comparar los resultados de un análisis de contenido hecho manualmente por analistas especializados con uno asistido por el programa computacional *SPSS Text Analytics for Surveys*
Objetivos específicos ✓ Contrastar los resultados en relación con los métodos manuales establecidos. ✓ Aplicar la Entrevista de Componentes Émicos del Bienestar. ✓ Aplicar la técnica de análisis de contenido heurístico y el análisis de contenido utilizando el programa *SPSS Text Analytics for Surveys 4.0.* ✓ Digitalizar el texto de cada cuestionario e importarlos al programa, el cual utiliza una combinación de técnicas lingüísticas y estadísticas automáticas para la extracción de términos y generación de categorías. ✓ Realizar conteo de frecuencias y obtener una lista de categorías para ser comparadas con el análisis de contenido manual.
Conclusiones: ✓ En líneas generales, revisando las extracciones y generando recursos afinados en el PLC se pueden obtener resultados similares al ACM. Sin embargo, el uso del PLC tiene las ventajas del menor tiempo y recursos consumidos. Adicionalmente, la confiabilidad aumenta al poder utilizarse una misma plantilla de generación de categorías para todo el análisis, en vez de utilizar varios investigadores o asistentes que realizan el análisis de contenido en paralelo, requiriendo una revisión cruzada. ✓ Otra ventaja del uso del PLC son las formas de visualización. En una ventana del programa, la respuesta original se muestra completa y únicamente se subraya la extracción (frase o término que luego será incluido en la categoría). En el ACM se suele leer la respuesta y reducirla a frases más cortas para insertarla en el archivo de trabajo, lo cual puede llevar a la pérdida de información o a caer en errores de lectura. Asimismo, el programa facilita la visualización de la categorización en el marco completo de la

respuesta, lo cual permite revisar la pertinencia de la inclusión de términos en cada categoría y favorece la comprensión de la información en su contexto de respuesta.

- ✓ El proceso repetido de extraer, revisar, refinar y volver a extraer potencializa la categorización con el PLC. A mayor uso del programa se van perfeccionando los recursos, mediante la adición de términos a la biblioteca y la creación de plantillas con fórmulas de categorización refinadas. Resulta una ventaja el hecho de que todos estos recursos puedan ser luego reutilizados y compartidos entre investigadores.
- ✓ Finalmente, el PLC permite diseñar investigaciones que eran poco factibles en el pasado. Los estudios de base son fundamentales para el desarrollo de instrumentos psicométricos sólidos cuyos ítems representen con precisión la variedad y amplitud de las variaciones en una población. Esto ha estado limitado por la complejidad y costos del análisis de contenido de muestras grandes. Con los PLC, resulta factible y económico realizar estudios de base con 500, 1500 o más participantes, abriendo el camino para estudios basados en entrevistas con alternativas de respuesta abierta representativos a nivel país. Estimamos que un análisis de contenido para 500 participantes utilizando los PLC debe tomar unas nueve horas. Una vez afinadas las bibliotecas y las plantillas, la diferencia para analizar 1500 participantes sería marginal. Este tipo de estudios permitiría una nueva generación de instrumentos psicométricos con ítems realmente representativos de la complejidad de una población.
- ✓ Existe un uso amplio de estudios basados en grupos focales o focus groups en la fase cualitativa, seguidos de una fase cuantitativa orientada en los resultados de los grupos focales. Sin embargo, es conocido que en los grupos focales prevalece el consenso grupal en contra del registro adecuado de la varianza individual, teniendo como ventaja un bajo costo. En este nuevo escenario, los costos de una serie de grupos focales seguidos de una fase cuantitativa serían menores a los estudios de entrevistas abiertas con muestras representativas. Adicionalmente, estos últimos tendrían la ventaja de que no se haría una inferencia desde los resultados cualitativos hacia la fase cualitativa, sino que se analizaría directamente la fase cualitativa aplicada a una muestra representativa. Esto tendría enormes implicaciones en los estudios aplicados en el ámbito de la investigación de mercados, los estudios organizacionales de base, así como en los diagnósticos para programas de desarrollo social.

✓ En conclusión, los PLC llegaron a su momento de madurez, existiendo programas como el Text Analytics for Surveys que ofrecen a un bajo costo, soluciones amigables y consistentes en una amplia variedad de lenguajes. Ofrece, con un debido manejo, un ahorro sustancial de tiempo, recursos, así como un incremento importante en la confiabilidad y de la validez. Abre puertas a una nueva generación de estudios que no estén restringidos por los costos y confiablidad de análisis de contenido de respuestas abiertas en muestras grandes. Ofrece un promisorio escenario para la investigación pura y aplicada. Las ciencias sociales se han sustentado con una sólida estructura cuantitativa que reposa en una deleznable estructura de estudios de base, sustentados en la cáscara de huevo de los grupos focales y los cimientos de entrevistas abiertas tan profundos como su tamaño muestral. En un futuro cercano, quizá seamos testigos de una importante reconstrucción de teorías y diagnósticos a través de estudios de base de una nueva generación. (2014)
Aporte al trabajo de investigación que se está realizando El proyecto demuestra que el análisis de contenido es una técnica que convierte las respuestas abiertas de entrevistas en categorías una solución automatizando el proceso, reduciendo drásticamente el tiempo requerido y mejorando la confiabilidad. Este proceso es de gran utilidad dado que define las categorías de un estudio sobre la base de la percepción de la muestra, evitando la imposición de categorías creadas por el investigador comparando un análisis de contenido hecho por expertos manualmente con el análisis del programa *SPSS Text Analytics for Surveys* (TA y abre las posibilidades para análisis cualitativos con muestras grandes.

2.1.2 Nacionales

Título: El análisis lingüístico a través de un corpus de entrevista oral
Autores: Álvarez Calderón, Cristian Giovanny Martínez Campos, Faiver Albeiro Navarro Sierra, Paula Andrea Alcántara
Universidad: Universidad Cooperativa de Colombia, Facultad de Ciencias Sociales Bucaramanga
Año: 2019

Objetivo general: Demostrar la pertinencia del empleo de entrevistas orales para determinar la expresión de anterioridad al momento del habla a través del análisis lingüístico.

Objetivos Específicos

- ✓ Sistematizar los referentes teóricos acerca la lingüística del corpus, la entrevista y el análisis lingüístico.
- ✓ Analizar el corpus de la lengua oral a través de la metodología establecida para el estudio de la temporalidad.
- ✓ Determinar las características fundamentales de la expresión de la anterioridad al momento del habla en la variedad colombiana del español.

Conclusiones

- ✓ Se sistematizaron variados documentos teóricos sobre la lingüística del corpus, la entrevista y el análisis lingüístico ya que ellos son la caracterización necesaria para trabajar con la lingüística del corpus. Ello se tomó como base inicial del proyecto, ya que son tres conceptos principales que hicieron posible la justificación del trabajo realizado con el fin de obtener un resultado general de ello.
- ✓ Estos tres aspectos resaltados en el proyecto inician con la lingüística del corpus, allí se trabajan conjuntos de datos electrónicos para analizar los registros de entrevistas orales identificando a su vez la lexicografía, la terminología y los estudios sociolingüísticos y multiculturales de la variedad del español de Santander. Tomando como base el corpus, se utiliza la entrevista oral semiestructurada ya que permite realizar preguntas abiertas y da la oportunidad a los participantes (entrevistador- entrevistado) de tener una conversación amena y en la que se pueda suficientes datos dados por las personas entrevistadas, además esta clase de entrevista es más acorde que la entrevista escrita porque el entrevistado tiene la oportunidad de dar su punto de vista de manera más cómoda, espontánea y abierta, lo cual lleva a un mejor análisis de los hablado. Después de la entrevista llega el momento del analizar lo realizado y es allí donde entra el análisis lingüístico la cual se compone de unas ramas que se comprenden dentro de un corpus de entrevista oral, como lo es, la fonética y fonología, la semántica y la pragmática, la sintaxis, la morfología, la metalingüística, la sociolingüística y donde la gramática también es necesaria para poder identificar el uso del momento del habla, permitiéndose el análisis

lingüístico desde la entrevista como objeto de análisis que infiere en el uso de las anterioridades en un determinado sintagma, pero cuando se transcribe a la escritura, esta se afianza más en el significado y uso del lenguaje como estudio de un objeto social, la relevancia en la inmersión social, cultural, familiar, personal y así, poder analizar, llegar a entender las diferencias de un contexto a otro y el uso pertinente de la lengua.

- ✓ A partir de los documentos teóricos consultados se realiza un análisis lingüístico utilizando el enfoque onomasiológico, que parte de la idea a la palabra, en 26 entrevistas, desarrolladas y analizadas desde una matriz, la cual es una herramienta eficaz para llevar a cabo un estudio de estos. Finalmente, para obtener resultados parciales, se identifican las variedades de expresión del español colombiano al momento del habla y se identifican características específicas de la expresión de anterioridad, el uso de los tiempos verbales que expresan pasado.
- ✓ Esta investigación fomenta un acercamiento no solo de lo semasiológico, estudia la relación que va desde el objeto a la palabra, esto significa que va sumergiéndose en el contenido onomasiológico, ya que analiza la anterioridad al momento del habla y sus formas de expresión en el español de Colombia desde la perspectiva de la lingüística cognitiva y de la teoría de los campos semánticos-pragmáticos y funcionales. (2019)

Aporte al trabajo de investigación que se está realizando

El desarrollo de este trabajo aporta una visión del análisis lingüístico del corpus a través de la entrevista oral a partir de la técnica de vaciado en la matriz, demostrando que los verbos y otros elementos aportan información temporal, además, no sólo los circunstantes temporales aportan sino aspectuales, actantes y de otros tipos de circunstantes.

2.1.3 Regionales

Al desarrollar la búsqueda y revisión sistemática y exhaustiva de literatura, no se encontró evidencia documental de investigaciones en las bases de datos de las principales universidades de la ciudad de San Juan de Pasto que coincidan con los propósitos, objetivos o fines de esta propuesta.

2.2 Marco Teóricas

En el siguiente capítulo se abordará temas esenciales dentro del proyecto de investigación, intentando dar una visión general de todos los aspectos teóricos que se relacionan con el proyecto, así como describir brevemente conocimientos relevantes, problemáticas y las técnicas relacionadas.

2.2.1 Hermenéutica

> *"La hermenéutica también nos sugiere y, sin duda, antes que toda otra consideración, un posicionamiento distinto con respecto a la realidad: aquel de las significaciones latentes. Se trata de adoptar una actitud distinta, de empatía profunda con el texto, con lo que allí se ha expresado a través del lenguaje. No se trata de suprimir o de intentar inhibir su propia subjetividad (con sus implícitos prejuicios), sino de asumirla. En otras palabras, la búsqueda de sentido en los documentos sometidos a análisis se ve afectada por un doble coeficiente de incertidumbre: la interpretación es relativa al investigador, así como al autor de los textos en cuestión"* (Baeza, 2002)

La hermenéutica fue originalmente una serie de técnicas para interpretar textos escritos, específicamente textos bíblicos con el propósito de descubrir y reconstruir lo que la gente piensa que el texto contiene como mensaje de Dios; el término se refiere a Hermes, quien era el mensajero de los dioses griegos y su propio Dios elocuente y astuto, convirtiendo la hermenéutica en un método de interpretación de texto qué hoy por hoy ya no se limita a obras religiosas; por su parte el teólogo alemán Friedrich Daniel Ernst Schleiermacher (1768-1834), impone un giro decisivo en su juicio de la teoría, ya que propuso la sistematización de la hermenéutica general como arte del comprender mismo, que sirviera de base a las teorías y metodologías para la interpretación de textos e intentó construir el sujeto del entendimiento en la expresión, el pensamiento y el dominio emocional del autor sin tratar de elaborar las reglas, pero también salvó la dimensión objetiva y creyó que se debe tomar en consideración la construcción del contexto del autor, es decir, pasar a la posición del autor y quitar la exclusividad subjetiva de la interpretación (Schleiermacher, 1829). Con Schleiermacher, nace según Dilthey (1944), la perspectiva de una teoría general de la ciencia y del arte de la interpretación.

Luego, la hermenéutica combina el texto y el lector en un proceso permanente de apertura y reconocimiento, tratando de ser objetivados para ser verificados; en este sentido, la hermenéutica, o más precisamente, las personas que la utilizan deben intentar realizar ejercicios de explicación contextual para comprender el texto. Este proceso implica desarrollar la comprensibilidad de las palabras contenidas en el texto, en gran medida se trata de traspasar los límites contenidos en la "física de las palabras" para capturar los significados expresados en papel; en este sentido, cobra relevancia el planteamiento de Ricoeur (1998), respecto a la necesidad de apreciar el análisis hermenéutico desde una concepción dialéctica: "La noción de acontecimiento de habla no está cancelada, más bien está sometida a una serie de polaridades dialécticas resumidas bajo el titulo doble de acontecimiento y sentido / significado y referencia. Estas polaridades dialécticas nos permiten anticipar que los conceptos de intención y dialogo no han de ser excluidos de la hermenéutica, sino más bien han de ser liberados de la unilateralidad de un concepto no dialéctico del discurso" (p. 19)

Así estamos en presencia de una doble posibilidad de interpretación, por cuanto el sentido puede ser captado desde lo que se quiso decir, específicamente la intencionalidad contenida en el discurso; y por otra parte, desde lo que realmente significa la oración esto es, en consideración a los elementos gramaticales y de vocabulario dispuestos en ella.

2.2.2 El concepto de corpus

Tomando como referencia el concepto corpus del lenguaje oral o la lingüística del corpus que es una muestra que es la encargada de emplear mediante varias herramientas la revisión y la organización de búsquedas, resultados relacionados con análisis lingüísticos y registrados previamente por escrito o de manera oral. Algunas de las áreas o campos que normalmente son consideradas en este tipo de análisis lingüísticos son “la lexicografía, la terminología, la traducción o los estudios sociolingüísticos y multiculturales”. (Garrote Salazar , 2010) Y es así como se puede encontrar a la ingeniería lingüística, área en la cual el corpus hace un papel importante y del cual se pueden derivar diferentes informaciones, dependiendo de los detalles y los objetivos iniciales, por ejemplo, se puede llegar a utilizar en uno o más de los niveles lingüísticos (fonológico, morfológico, sintáctico y textual), hasta el análisis de tiempos verbales, anterioridades, circunstantes y situaciones discursivas, sentimientos entre otros. En la actualidad el concepto de corpus ha cambiado mucho con respecto al que manejaban los primeros lingüistas;

hoy en día se considera que los corpus deben cumplir los siguientes requisitos. (McEnery & Wilson, Corpus Linguistics, 1996)

2.2.2.1 Textos en formato electrónico: un corpus, para ser una herramienta útil al lingüista, debe estar informatizado, es decir, los textos de que consta tienen que estar en formato electrónico (corpus informatizado o automatizado). El hecho de que para los primeros corpus no se pudiera disponer de ordenadores motivó la crítica de las pseudotécnicas: el procesamiento de los datos debía efectuarse de forma manual, con los errores y problemas que eso ocasionaba. Sin embargo, el empleo del ordenador permite automatizar tareas tales como: Búsqueda de información ya que permite localizar de forma rápida una palabra, una secuencia de palabras o incluso una categoría gramatical en décimas de segundo, recuperación de información dado que permite obtener todos los casos de una palabra, secuencia de palabras, etc. registrados en el corpus, normalmente con su contexto inmediato anterior y posterior (concordancia), cómputo de la frecuencia de aparición de una palabra, secuencia de palabras, etc., clasificación de los datos contenidos en el corpus según diferentes criterios: orden alfabético, frecuencia de aparición, autor, procedencia geográfica, tema, medio de publicación, etc..

2.2.2.2 Autenticidad de los datos: los textos recogidos en el corpus deben ser muestras reales de uso de la lengua objeto de estudio, y a partir de ellas se construyen (o verifican) de forma empírica las teorías que tratan de explicar el funcionamiento de la lengua o las aplicaciones computacionales, criterios de selección puesto que los textos que forman parte del corpus deben haber sido elegidos de acuerdo con unos determinados criterios lingüísticos y/o extralingüísticos para la finalidad concreta que persiga el corpus.

2.2.2.3 Representatividad: la selección de los textos, además de tener unos criterios adecuados, debe responder a parámetros estadísticos que garanticen que los textos “representan” la variedad de lengua objeto de estudio (muestra representativa) recurriendo a la selección, según criterios estadísticos, de textos de diversos géneros, tipologías, temas, medios de publicación, etc.

2.2.2.4 Tamaño: el tamaño solo es importante en la medida en que así lo exija la finalidad del corpus, por lo general, los corpus constan de un tamaño finito, que se suele medir en millones de

palabras (o formas) y que se fija antes de empezar la recogida de los textos (p. ej. un millón de palabras); una vez alcanzado ese número, se da por terminada la recopilación del corpus, que no es más que el primer paso de todo el proceso, sin embargo, también existen corpus abiertos o monitor, como el del proyecto COBUILD[1] dirigido por J. Sinclair en la Universidad de Birmingham, de especial interés para la lexicografía.

A continuación, se recogen algunas definiciones de corpus que ilustran estas características:

- ✓ Una colección de textos que se supone que son representativos de un idioma, dialecto u otro subconjunto de un idioma dado, que se utilizará para el análisis lingüístico. (FRANCIS, W. N., 1992)
- ✓ Una colección de piezas de lenguaje que se seleccionan y ordenan de acuerdo con criterios lingüísticos explícitos para ser utilizados como muestra del lenguaje. (Sinclair, 1996)
- ✓ Un cuerpo de tamaño finito de textos legibles por máquina muestreados con el fin de ser lo más representativo de la variedad de idiomas en consideración. (McEnery & Wilson, Corpus Linguistics, 1996, pág. 24).
- ✓ Un corpus es una muestra de una lengua que habitualmente se construye a partir de una selección de textos realizada según unos determinados criterios y un determinado objeto. (Martí Antonín & Castellón Masalles, 2000, pág. 151)
- ✓ El término corpus solo debe aplicarse correctamente a una colección bien organizada de datos, recopilados dentro de los límites de un marco de muestreo diseñado para permitir la exploración de una determinada característica lingüística (o conjunto de características) a través de los datos recopilados. (McEnery, "Corpus Linguistics", en R. Mitkov (ed.), 2003).
- ✓ Un corpus es un conjunto de textos de lenguaje natural e irrestricto, almacenados en un formato electrónico homogéneo, y seleccionados y ordenados, de acuerdo con criterios explícitos, para ser utilizados como modelo de un estado o nivel de lengua determinado, en estudios o aplicaciones relacionados en mayor o menor medida con el análisis lingüístico (Santalla del Río, 2005)

[1] Acrónimo de *Collins Birmingham University International Language Database* , es un centro de investigación británico establecido en la Universidad de Birmingham en 1980 y financiado por los editores de Collins . http://www.collins.co.uk/books.aspx?group=140

✓ Por el contrario, el término corpus, tal como se utiliza en la lingüística moderna, se puede definir mejor como una colección de textos de muestra, escritos o hablados, en forma legible por máquina que puede anotarse con diversas formas de información lingüística. (McEnery, T., Xiao, R. y Tono, Y., 2006)

Estos criterios y definiciones permiten discriminar los corpus, en el sentido que se maneja en la lingüística de corpus, de otras colecciones de textos electrónicos. (Torruella & Llisterri, 1999):

- ✓ **Archivo (o colección) informatizado**: se trata de un simple conjunto de textos electrónicos sin estructurar. El único criterio que prevalece a la hora de conformarlo es la disponibilidad de los textos.
- ✓ **Biblioteca de textos electrónicos**: se trata de un conjunto de textos electrónicos recogidos sin seguir criterios lingüísticos, pero guardados en un formato estándar.

2.2.2.1. Clasificación de los corpus

En este apartado se propone una clasificación de los diferentes tipos de corpus de acuerdo con su finalidad y objetivo basada en autores como (Sinclair, 1996) o (Torruella & Llisterri, 1999) quienes han propuesto clasificaciones en función del objetivo que se persigue con el corpus y su finalidad*:*

Tabla 1 Clasificación de los corpus

Clasificación de los Corpus			
1. Según la modalidad de la lengua			
Los corpus textuales o escritos	**Los corpus orales**	**Los corpus grabaciones**	**Los corpus mixtos**
Conformados exclusivamente por muestras de lengua escrita. Ejemplo *Corpus Textual Informatitzat de la Llengua Catalana (CTILC).*[2]	Recogen únicamente muestras de lengua hablada, que pueden ser transcripciones ortográficas de grabaciones para obtener una representación simbólica de una muestra natural de habla. Esta transcripción constituye el punto de partida para el tratamiento	Los corpus orales orientados hacia la descripción fonética de las lenguas suelen consistir en inventarios de sistemas fonéticos y fonológicos de las lenguas del mundo a modo de bases de datos de sonidos o en grabaciones realizadas en condiciones óptimas de segmentos	Combinan ambas modalidades de lengua, aunque siempre favorecen la lengua escrita, ya que su obtención es menos costosa que la de la lengua oral que, además, requiere un proceso posterior de transcripción de

[2] *https://ctilc.iec.cat/scripts/*

	posterior del corpus (añadir marcas sobre categorías gramaticales, extraer índices de frecuencia, determinar sentimientos etc.) y para efectuar diferentes análisis lingüísticos: sociolingüísticos, discursivos, etc. **Ejemplo:** *The Bergen Corpus of London Teenage Language (COLT)*[3] que es un corpus de medio millón de palabras conformado por las transcripciones ortográficas de conversaciones espontáneas. Su objetivo fundamental es dar cuenta de una variedad de lengua de los adolescentes de Londres y, por tanto, servir como punto de referencia para estudios de índole lingüística Corpus Oral de Lenguaje Adolescente (COLA)[4] Corpus de Conversación Coloquial del Grupo *Val.Es.Co*[5] Corpus Oral de Referencia de la Lengua Española Contemporánea (CORLEC).[6] Proyecto PRESEA[7] para la creación de un corpus representativo de las variedades geográficas y sociales del español. El proyecto, dirigido por F.	aislados, frases aisladas o textos leídos. **Ejemplo:** *Corpus Albayzín* (CASACUBERTA, F. et al., 1992) el cual es una gran base de datos oral desarrollada en España, entre 1992 y 1998, por un consorcio de grupos de investigación en tecnología del habla coordinado por la Universidad Politécnica de Cataluña. Además de los objetivos relacionados directamente con la síntesis y el reconocimiento del habla con vistas al desarrollo de estudios fonéticos sobre la variabilidad inter- e intra-locutor, la variabilidad contextual y la variabilidad condicionada por las condiciones ambientales. Proyecto EUROM (*vid. Chan et al. 1995*), es una base de datos oral multilingüe, en la que las grabaciones se llevaron a cabo bajo las mismas condiciones, con el mismo número de sujetos y un corpus equivalente para once lenguas del entorno.	grabaciones. **Ejemplo:** Corpus de Referencia del Español Actual (CREA) *British National Corpus (BNC)* que pertenecen a este tipo de corpus ya que el 90% de sus textos son escritos y el 10% restante, orales.

[3] http://korpus.uib.no/icame/colt/
[4] https://cola.w.uib.no/
[5] https://www.uv.es/corpusvalesco/corpus.html
[6] https://ila.uca.es/corpus-oral-de-referencia-de-la-lengua-espanola-contemporanea-corlec/
[7] https://preseea.linguas.net/

	Marcos Marín en el Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid, se realizó entre 1991 y 1992.		
2. Según el número de lenguas			
Los corpus monolingües		**Los corpus bilingües o multilingües**	
Compuestos por textos en una sola lengua con el objetivo de dar cuenta de dicha lengua o variedad lingüística (o de un subconjunto de la misma). Ejemplo: CREA (para el español) CORGA (para el gallego)		Formados por textos de dos (bilingües) o más lenguas (multilingües) sin que, en principio, sean traducciones unos de otros y sin compartir criterios de selección. **Ejemplo:** *International Corpus of English (ICE),* un corpus en el que desde 1990 se están recopilando materiales escritos y orales posteriores a 1989 pertenecientes a diferentes variedades del inglés del mundo. Es un tipo de Corpus comparables *("paired texts")* C-Oral-Rom, corpus multilingüe de habla espontánea de cuatro lenguas romances (italiano, francés, portugués y español). Corpus literario TECTRA inglés-galego (1.476.020 palabras) Corpus literario FEGA francés-galego (1.648.272 palabras) Corpus xurídico LEGA galego-español (6.582.415 palabras) Corpus UNESCO inglés-galego-francés-español de divulgación científica (3.724.620 palabras) Corpus LOGALIZA de localización de software inglés-galego (3.526.850 palabras) Corpus CONSUMER español-galego-catalán-euskara de información sobre consumo (5.586.431 palabras)	
3. Según la cantidad, proporción y distribución de los tipos de textos			
Corpus grandes	**Corpus equilibrados**	**Corpus piramidales**	**Corpus léxicos**
No tienen un límite de palabras o este es muy elevado en comparación con otros tipos de corpus; no suelen atender a cuestiones de equilibrio o de representatividad. Cada vez es mayor la tendencia	Recogen la misma proporción de diferentes tipos de textos.	Contienen textos distribuidos en estratos o niveles, de tal forma que un nivel consta de pocas variedades temáticas, pero con muchos textos para cada una; un segundo nivel, de textos más	Recogen fragmentos de textos muy pequeños y de longitud constante en cada documento. Era lo habitual en los primeros corpus, debido a las

al aumento de volumen gracias a los medios y facilidades técnicas disponibles; no obstante, en la actualidad existen corpus de gran tamaño diseñados con criterios que garantizan la representatividad de los datos.		variados temáticamente, pero con menos cantidad de cada uno; etc.	limitaciones de tamaño que los medios técnicos de la época imponían. Hoy en día han vuelto a cobrar importancia debido a lo cuidado de su diseño.
4. Según los límites establecidos			
Los corpus cerrados		**Los corpus abiertos o corpus monitor**	
Constan de un número finito de palabras, que se establece de forma previa a la recopilación del corpus. Una vez alcanzado ese número, el corpus se da por finalizado, sin añadir más material posteriormente. **Ejemplo:** *Corpus Brown*		Son corpus dinámicos, que se mantienen constante crecimiento, mediante introducción periódica de nuevas cantidades de textos según unas proporciones definidas. Es un excelente material de estudio diacrónico para observar tendencias de uso, cambios de significado, frecuencias de distribución, etc. **Ejemplo:** *Bank of English*	
5. Según la especificidad de los textos			
Los corpus generales o de referencia	**Los corpus especializados**	**Los corpus genéricos**	**Corpus canónicos**
Este tipo de corpus intentan reflejar la lengua o variedad lingüística equilibradamente, tienen que ser suficientemente amplios para reflejar todas las variedades relevantes de una lengua y su vocabulario de forma que, se puedan tomar como base para la elaboración de gramáticas, diccionarios, tesauros, etc. **Ejemplo:** El corpus CREA.	Estos corpus almacenan textos que puedan aportar datos para la descripción de un tipo particular de lengua ("*sub-lenguaje*"). Con la meta de estudiar cómo funciona la lengua en cada una de esas áreas y extraer información útil para detectar neologismos, elaborar diccionarios y tesauros, estudiar la variación lingüística, etc. **Ejemplo:** Corpus Técnico do Galego (CTG) del Seminario de Lingüística Informática de la Universidad de Vigo contiene textos jurídico-administrativos, de informática y telecomunicaciones, de ecología y ciencias ambientales, de economía, de sociología y de medicina.	Recogen textos pertenecientes a un único género, su objetivo es caracterizar un género frente a otro. **Ejemplo:** *Corpus York-Helsinki Parsed* *Corpus of Old English Poetry*, contiene solo poesía.	Formados por todos los textos que configuran la obra completa de un autor.

	Corpus textual especializado plurilingüe, del Instituto Universitario de Lingüística Aplicada de la Universidad Pompeu Fabra, consta de textos en catalán, castellano, inglés, francés y alemán sobre economía, derecho, medio ambiente, medicina e informática.		

6. Según el periodo temporal que abarcan los textos.		
Los corpus periódicos o cronológicos	**Los corpus diacrónicos o históricos**	**Corpus sincrónicos**
Contienen textos de unos años determinados o de unas épocas concretas con el objeto de estudiar la lengua producida durante ese período. **Ejemplo:** Corpus Brown Recogen textos publicados exclusivamente en 1961 en Estados Unidos Corpus LOB Recogen textos publicados exclusivamente en 1961 en el Reino Unido	Recogen textos de diferentes etapas temporales sucesivas con el fin de poder observar evoluciones de la lengua en un período largo. **Ejemplo** CORDE Corpus del español, un corpus de cien millones de palabras recopilado por Mark Davis en la Universidad de Brigham Young contiene textos en español desde el siglo XIII hasta el XX.	Este tipo de corpus buscan permitir el estudio de una o más variedades lingüísticas en el momento presente, sin prestar atención a su evolución excepto en lo que se refiere a los cambios rápidos que ocurren en la actualidad. **Ejemplo** *Corpus of Contemporary American English*, de más de trescientos ochenta y cinco millones de palabras procedentes de textos de diferentes fuentes de los años 1990 a 2008.

7. Según el proceso al que se someta el corpus.					
Corpus simples	**Corpus verticales**:	**Corpus codificados o anotados**	**Corpus analizados morfológicamente *(“tagged”)***	**Corpus “*parentizados*”:**	**Corpus analizados (“*treebanks*”)**
Son corpus con textos guardados sin formato alguno y sin añadir ningún tipo de información adicional, como pueden ser códigos o anotaciones o codificación.	En este tipo de corpus se dispone en forma de columna las palabras de un texto ordenadas según criterios alfabéticos o de frecuencia. Las palabras se	Estos corpus están formados por textos a los que se les han añadido, de forma manual o automática, determinada información como datos bibliográficos o autor, el título, los capítulos, los párrafos, etc. (codificación); o,	En estos corpus los textos son anotados con información morfológica. Cada palabra del corpus tiene asociada una lista de sus posibles categorías morfosintácticas como el nombre o verbo.	Estos corpus tienen un proceso de análisis sintáctico superficial, nominal, verbal, etc. **Ejemplo** *Lancaster Parsed Corpus (LPC)*, representa un subconjunto	En este tipo de corpus los textos que lo conforman están procesados sintácticamente de manera completa. Cada oración del corpus ha sido analizada de forma exhaustiva

	consideran aisladament e, sin contexto	lo que es más interesante, a aspectos puramente lingüísticos, como la categoría gramatical, la estructura sintáctica, etc. (anotación).		del LOB de unas ciento cuarenta mil palabras que han sido analizadas sintácticament e.	**Ejemplo** Base de Datos Sintácticos del Español Actual (BDS) Corpus CESS-ECE para el español, el catalán y el euskera AnCora, para el español y el catalán.

Fuente: Esta investigación.

Vemos como la tipología de los corpus pueden ser clasificados por el fin que persigue como el estudio de la obra de un autor como Dickens o de la producción literaria de una época determinada como el Rococó, la descripción de una lengua en general como el español contemporáneo o de un sub - lenguaje o aspecto lingüístico concreto como el léxico jurídico o médico y en nuestro caso particular de estudio la detección de sentimientos para medir emociones negativas o positivas para un objetivo específico.

2.2.3 La lingüística del corpus

La lingüística del corpus es “un conjunto de textos almacenados en formato electrónico y agrupados con el fin de estudiar una lengua o una determinada variedad lingüística” (s.f., 2020) siendo una rama de la lingüística que basa sus investigaciones en datos obtenidos a partir del corpus, es decir de muestras reales de lenguaje; Rojo (2002) afirma que en la recolección de datos o materiales se diferencia la lingüística del corpus de cualquier otra manera de análisis del lenguaje, debido a que es posible que en ella, el lingüista pueda realizar “una búsqueda sobre un conjunto de textos y recibe, como resultado de una búsqueda ‘ciega’ llevada a cabo por la máquina, todos los casos a veces en cantidades realmente aplastantes que responden formalmente a lo que ha solicitado.” (Rojo., 2002) De esta forma, es posible comprobar el número que puede cubrir una sola investigación o desarrollo de corpus, para obtener y sustentar los fenómenos de lenguaje en las muestras de habla del corpus, pues utilizando este método se pueden considerar y

recolectar todos los datos del corpus, y también analizar el comportamiento de los fenómenos del lenguaje para obtener datos porcentuales, estadísticos o cualitativos; estos datos hacen una gran contribución a la interpretación de la gramática y la detección de emociones positivas, negativas o neutrales.

2.2.4 Desarrollo diseño y constitución del corpus

El desarrollo diseño y constitución del corpus se refiere al proceso mediante el cual se fijan los criterios que han de guiar el desarrollo del corpus, los pasos a seguir para su diseño y constitución son:

2.2.4.1 Selección del Tema. La selección de tema y estándares lingüísticos se basa en la distribución de palabras o rasgos gramaticales, y se centra en la diversidad lingüística del texto seleccionado o en la aparición de elementos diferenciadores, como la longitud de la oración, que produce un tipo de texto, para que un corpus sea representativo debe contener la mayor cantidad de palabras o rasgos lingüísticos más representativos de las variantes estudiadas. Al seleccionar el campo o dominio al que pertenece el texto que se selecciona según el contenido del texto, correspondiente al estándar interno en base a la descripción y clasificación del campo de conocimiento; para reducir la complejidad a una estructura jerárquica unidimensional es necesario dividirla en seis campos, agrupando una gran cantidad de áreas temáticas:

Considerando que el corpus lingüístico se crea a partir de las tesis de maestría de la universidad de Nariño con el fin de ser apoyo hermenéutico para las investigaciones de corte cualitativo, el tema que lo agrupa seria las ciencias de la educación, y, en articulación con las líneas de investigación se proponen los siguientes hipercampos para el corpus:

- Análisis curricular
- Arte, Artesanías y formación de región
- Comprensión, interpretación y producción de textos argumentativos
- Conocimiento y naturaleza en las comunidades rurales y urbanas en Nariño
- Creatividad cultural en colectivos locales
- Creatividad Social
- Currículo y Universidad
- Currículos pertinentes

- Diseño, Comunicación y Procesos Interactivos
- Educación Ciencia y Tecnología para el Desarrollo Sostenible
- Educación virtual
- Enseñanza de las Ciencias
- Historia curricular
- Historia de la cultura jurídica
- Historia de la educación en América Latina
- Historia Regional
- Historia Regional
- Historia regional y procesos de formación de nación
- Innovaciones Pedagógicas
- Literatura y región
- Pedagogía
- Pedagogía Social
- Tecnologías de la Información y la Comunicación para la Educación

2.2.4.2 Establecimiento de Bases Conceptuales de Selección del Corpus. La elección de material es la base conceptual que alimentara el corpus, esta depende del objetivo del corpus. Para nuestro caso particular se realizará de acuerdo con la muestra de trabajos de grado presentados por los estudiantes de la Maestría en Educación de la universidad de Nariño entre los años 2017 y 2021.

2.2.4.3 Etiquetado de los Textos. Corresponde a la fase de codificación y programación del material de entrenamiento de análisis lingüístico para el corpus el cual en nuestro caso de estudio se codificará a partir del ejercicio hermenéutico de una muestra de las tesis de Maestría en Educación desde el año 2017 hasta la actualidad**.**

2.2.4.4 Procesamiento Informático de los Textos. Esta fase se realiza un procesamiento informático de los textos etiquetados y codificados para la materialización e implementación del soporte lógico y físico (hardware y software) del corpus**.**

2.2.4.5 Análisis Lingüístico. El análisis lingüístico es el objetivo de un corpus ya sea de tipo cualitativos donde se estudian las características de una lengua o algún fenómeno ocurrido en esta o de tipo cuantitativos donde se analiza todo lo competente a cifras numéricas, frecuencias de aparición, etc. mediante criterios lingüísticos pertinentes no solo para contener el texto en sí mismo, sino que además proporcione información que facilite su análisis.

2.2.4.6 Análisis de Sentimientos. El análisis de sentimientos analiza el vocabulario de un texto para determinar sus cargas emocionales, su propósito es determinar los sentimientos o cargas emocionales de un texto, haciendo uso de lexicones para procesar y reconocer dichos sentimientos. En este sentido hay dos técnicas aproximadas en el contexto del análisis de textos, la minería de sentimientos relacionados y opiniones tanto positivas como negativas, la cual se centra en determinar la polaridad del texto y el análisis de sentimientos que se centra en identificar sentimientos, pero dado que el análisis de sentimientos es usado para la detección de polaridad se suele considerar que ambas técnicas son una misma cosa.

El análisis de sentimientos tiene por objetivo rescatar la polaridad de la opinión del autor de un texto, llevando el análisis de textos hacia un alcance mayor, la evaluación de intenciones y orientación de un texto, considerando tanto elementos explícitos como implícitos del lenguaje, por esto el análisis de textos como tarea del procesamiento de lenguaje natural (PLN) en un corpus lingüístico que apoye el proceso hermenéutico en una investigación se centra en la clasificación de un conjunto de documentos en categorías, es decir la categorización automática de textos, permitiendo el desarrollo de algoritmos para ordenar y recuperar información de manera inteligente.

2.2.4.6.1 Problemática del Lenguaje Natural. El principal problema al que se enfrenta un corpus lingüístico en la tarea de detectar sentimientos es la ambigüedad del lenguaje natural, la cual se presenta como distintos fenómenos a distintos niveles del lenguaje.

- ✓ **Homonimia:** se producen cuando se escriben palabras con diferentes significados, o cuando la pronunciación es la misma en el caso del lenguaje hablado. En el primer caso, estamos hablando de homógrafas, y en el segundo caso, estamos hablando de homófonas. Para nuestro caso de lenguaje natural escrito, esto significa que el corpus no puede reconocer palabras con total certeza solo comparando sus formas, por lo que

el sistema debe proporcionar algún mecanismo de desambiguación que pueda distinguir entre las dos situaciones.

- ✓ **Polisemia:** La polisemia ocurre cuando una palabra tiene múltiples significados. En este caso, diferentes significados provienen de un origen común y se crean debido a la evolución de su significado.
- ✓ **Anáforas y elipsis**: Las elipsis es la supresión de elementos en la estructura sintáctica, porque estos pueden inferirse del contexto. En una manera similar, las anáforas son un elemento deíctico cuyo significado depende también del contexto, donde su uso resulta útil en el lenguaje habitual para que la comunicación sea más fluida, pero que añade una gran dificultad a la hora de su procesado.
- ✓ **Sintaxis no normalizada** es una dificultad que trae el análisis del lenguaje natural y la detección de sentimientos el que su estructura no está estandarizada, es decir, al usar el idioma español, no usa una sintaxis específica y claramente definida dado que la misma idea se puede expresar de diferentes formas como, por ejemplo:

  - Me encanta el café.
  - El café me encanta.

En las dos oraciones se expresa la misma idea utilizando un orden diferente en las palabras que las componen.

***2.2.4.6.2 Procesamiento del Lenguaje Natural*.** A continuación, se hará una descripción general de las técnicas de procesamiento del lenguaje natural con las que contaremos para realizar el corpus:

- ✓ ***Tokenización***: La primera técnica que se realiza en el texto que se va a procesar mediante la tecnología PLN es la *tokenización*, incluye la división en "bloques", llamados tokens, es decir, el flujo de caracteres que componen el texto. Esto sucede en todos los niveles, básicamente en frases y palabras. La separación del texto permite un análisis independiente de cada elemento, y el establecimiento de la relación entre las etiquetas de resultado suele basarse en el uso de puntuación en el texto, como puntos, comas, espacios y signos de interrogación, uso de emoticones para dividir el texto en bloques independientes con sentido y significado independiente.

- ✓ **Lematización**: El proceso de lematización implica asignar a cada palabra su lema como su forma canónica. Usar la forma canónica para representar cada palabra en el texto permite crear reglas simples para afectar un conjunto de palabras sin tener que escribir todas sus variantes. Por ejemplo, la forma inflexiva de un adjetivo son sus diferentes cambios de género y cantidad, muchos o muchas, al igual que los verbos se asignan como infinitivos de la forma canónica.
- ✓ ***Part-of-speech tagging (POS)****:* El proceso de *Part-of-speech tagging*, abreviado POS, se corresponde con el análisis morfosintáctico del texto. Es el encargado de asignar a cada token del texto analizado su categoría morfosintáctica. De esta manera, luego de que el texto haya pasado este procesamiento, a cada etiqueta se le asignará una etiqueta, la cual indicará todos los aspectos relacionados con su gramática: categoría gramatical, género, número, persona, y otros aspectos relevantes relacionados con la palabra específica que se está analizando.
- ✓ **Desambiguación**: Los procesos de desambiguación resuelven los problemas del procesamiento del lenguaje natural por la naturaleza polisémica del lenguaje mediante restricciones morfosintácticas, las cuales desambiguan el sentido concreto de una palabra en función de las categorías gramaticales que se usen en la construcción de la oración como por ejemplo “Pasto” como equipo de futbol frente a “Pasto” como ciudad.
- ✓ **Análisis sintáctico**: El análisis sintáctico añade una descripción funcional a cada parte de la oración, estableciendo además un árbol sintáctico con distintos niveles de segmentación en los que se establecen relaciones gramaticales del tipo sujeto, complemento del verbo, etc.

2.2.4.6.3 Machine Learning. La Lingüística de corpus genera una serie de métodos de investigación, tratando de trazar un camino de datos a la teoría, mediante el *Machine Learning* se permite entrenar modelos e inferir patrones del lenguaje mediante las 3ª; (Wallis, S. & Nelson G., 2001) introdujeron por primera vez lo que ellos llamaron la perspectiva de las tres A (3A perspective): anotación, abstracción y análisis para que los modelos de aprendizaje automático procesen el lenguaje con cada vez más precisión, entrenar modelos de aprendizaje supervisado y detectar determinados fenómenos lingüísticos.

Uno de los elementos fundamentales para el desarrollo de la herramienta de Corpus desarrollada es el Word2Vec Mikolov et al (2013) basada en una idea introducida por Firth en 1957: “el significado de las palabras está asociado a las palabras que le acompañan (contexto)” es decir que busca la compresión y procesamiento asistidos por ordenador de información expresada en lenguaje humano mediante el análisis del leguaje humano y su interpretación, dando significado para que pueda ser utilizado de manera práctica. Mediante la utilización del Procesamiento natural del lenguaje PNL se puede hacer tareas como resumen automático de textos, traducción de idiomas, extracción de relaciones, análisis de sentimiento, reconocimiento del habla y clasificación de artículos por temáticas, armar diversos modelos con el lenguaje, crear estructuras y con ellas alimentar algoritmos de *Machine Learning*.

En el desarrollo de la herramienta se usó la técnica de *Machine Learning* combinada con el Procesamiento natural del lenguaje PNL, como una disciplina del ámbito de la Inteligencia Artificial creando un algoritmo sistemático que aprende automáticamente, es decir, un proceso mediante el cual se identifica patrones en una gran cantidad de datos a través de una máquina que aprende, que en este caso es un algoritmo encargado de revisar los datos y puede detectar sentimientos positivos, negativos y neutros así como definir y clasificar categorías de manera automática. Esta simbiosis entre el *Natural Language Processing* y el *Machine Learning* es lo que se conoce como Inteligencia Cognitiva.

El punto de partida del proyecto es el corpus, el cual es un conjunto de textos, ordenados, que sirven de base para cualquier análisis lingüístico, posteriormente se desarrolló una anotación sistemática y exhaustiva, que convierte el conjunto de textos en un corpus anotado, realizando sobre el texto un etiquetado preciso de cada término positivo, negativo o neutro así como su categorización, imprescindible para que la herramienta pueda comenzar a actuar sobre esa información y entrenar el modelo de aprendizaje que detecten determinados fenómenos lingüísticos.

Como afirma Tom Mitchell, el aprendizaje de máquina es un área que estudia cómo construir programas de computadoras que mejoren su desempeño en alguna tarea gracias a la experiencia. (1997)Este aprendizaje está basado en ideas de diversas disciplinas, como inteligencia artificial, estadística y probabilidad, teoría de la información, psicología y neurobiología, teoría de control y complejidad computacional. (Nilsson, N. J., 1998) Para el entrenamiento de la herramienta del corpus lingüístico desarrollado en el presente proyecto para utilizar el abordaje de aprendizaje, se

consideraron una serie de decisiones que incluyen la selección de los documentos para el entrenamiento y alimentación de la herramienta correspondientes a las tesis desarrolladas en la Maestría de Educación de la Universidad de Nariño correspondientes a una muestra representativa obtenida del total de trabajos recibidos desde su creación hasta el presente con los cuales se obtiene el dataset para entrenar al modelo de *Machine Learning*, de acuerdo con el tipo de entrenamiento que requiere la herramienta para el desarrollo de la función objetiva a ser aprendida, su representación y el algoritmo para aprender esa función como un proceso inductivo que automáticamente construye un clasificador que procesa de diversas maneras el lenguaje en sus componentes: gramática, sintaxis e intenta crear estructuras de apoyo que sirven como entradas para aplicar regresión lineal, regresión logística, *naïve bayes*, árbol de decisión o redes neuronales, aprendiendo a partir de un conjunto de documentos preclasificados para darles un análisis morfológico o léxico, que es el análisis interno de las palabras que forman oraciones para extraer lemas, rasgos flexivos y unidades léxica compuestas lo cual es indispensable para la información básica de la herramienta desarrollada determinar la categoría sintáctica y significado léxico del material de entrenamiento de la herramienta de acuerdo con el modelo gramatical empleado lógico y estadístico, además se realiza un análisis semántico, lo cual proporciona la interpretación de las oraciones una vez eliminadas las ambigüedades morfosintácticas; acompañado de un análisis pragmático que permite analizar el contexto de uso a la interpretación final de la palabra incluyendo el tratamiento de las ironías y metáforas u otras formas de lenguaje figurado para concluir con una adecuada detección de sentimientos y categorías al usar la herramienta en investigaciones de tipo cualitativo.

Hsinchun Chen y Michael Chau (2004) mencionan los principales paradigmas del aprendizaje de máquina, que son el modelo probabilístico, el aprendizaje simbólico y la inducción de reglas, las redes neuronales, los algoritmos basados en evolución, el aprendizaje analítico y los métodos híbridos. En el desarrollo de la herramienta se utilizaron predominantemente el paradigma de red neuronal de aprendizaje de máquina para aprovechar mejor las ventajas que presenta ya que imita a las neuronas humanas; los conocimientos son representados por descripciones simbólicas; el conocimiento es aprendido y recordado por redes de neuronas artificiales interconectadas por sinapsis con pesos y unidades de umbral lógicas. (Hsinchun & Chau, 2004). Las ventajas de esta técnica sobre el enfoque de ingeniería del conocimiento o análisis de encuestas de forma manual son una muy buena efectividad, ahorros considerables en términos de tiempo, mano de obra

experta y una portabilidad sencilla a diferentes dominios abordando con éxito tareas consideradas extremadamente complicadas; en cierta forma, estos resultados nos acercan un poco al objetivo final del Procesamiento del Lenguaje Natural, que no es otro que: «las máquinas entiendan realmente el lenguaje».

2.3 Marco Legal

Las siguientes regulaciones son consideradas como las más importantes sobre propiedad intelectual en Colombia y en el ámbito internacional:

En Colombia, la Asamblea Nacional Constituyente, adoptó en la (Constitución política de Colombia, 1991), el texto del artículo 61, que expresa: "El Estado protegerá la propiedad intelectual por el tiempo y mediante las formalidades que establezca la ley".

- ✓ LEY 23 DE (Congreso de la República de Colombia. , 1982). Sobre derechos de autor. Congreso de la República de Colombia. Los autores de obras literarias, científicas y artísticas gozarán de protección para sus obras en la forma prescrita por la presente Ley y, en cuanto fuere compatible con ella, por el derecho común. También protege esta Ley a los intérpretes o ejecutantes, a los productores de programas y a los organismos de radiodifusión, en sus derechos conexos a los del autor.
- ✓ LEY 44 DE (Congreso de la República de Colombia, 1993) Personas que sean autores de obras protegidas por el Derecho de Autor, podrán disponer contractualmente de ellas con cualquiera entidad de derecho público.
- ✓ Ley 1951 de (Congreso de la República de Colombia, 2019) Por la cual se crea el Ministerio de Ciencia, Tecnología e Innovación, se fortalece el Sistema Nacional de Ciencia, Tecnología e Innovación y se dictan otras disposiciones sobre derechos de autor para tecnología.
- ✓ Decreto 393 del 08 de febrero de (Presidencia de la Republica de Colombia, 1991) Por el cual se dictan normas sobre asociación para actividades científicas y tecnológicas, proyectos de investigación y creación de tecnologías.
- ✓ CONVENCIÓN DE ROMA. Sobre la protección de los artistas intérpretes o ejecutantes, los productores de fonogramas y los organismos de radiodifusión. Hecho en Roma el 26 de octubre de 1961.
- ✓ CONVENIO DE BERNA. Para la Protección de las obras literarias y artísticas. Acta de París del 24 julio de 1971 y enmendado el 28 de septiembre de 1979.

- ✓ DECISIÓN 486. Por la cual se establece el marco legal de propiedad intelectual para los países Comunidad Andina de Naciones, CAN: Colombia, Venezuela, Ecuador, Perú y Bolivia. Septiembre de 2000.
- ✓ TRATADO DE LA OMPI SOBRE DERECHOS DE AUTOR. El tratado obliga a las partes contratantes a proveer remedios legales contra la anulación de las medidas tecnológicas que emplean los autores en el ejercicio de sus derechos y contra la remoción o alteración de información, como ciertos datos que identifican la obra de sus autores, que es necesaria para la administración de sus derechos. Adoptado en Ginebra el 20 de diciembre de 1996.
- ✓ TRATADO DE LA OMPI SOBRE INTERPRETACIÓN O EJECUCIÓN Y FONOGRAMAS. Este tratado se refiere a los derechos de propiedad intelectual de dos tipos de beneficiarios: intérpretes (actores, cantantes, músicos, etc.) y productores de fonogramas (las personas o entidades legales que toman la iniciativa y asumen la responsabilidad en relación con la grabación de esos sonidos). Adoptado en Ginebra el 20 de diciembre de 1996.
- ✓ GACETA OFICIAL DEL ACUERDO DE CARTAGENA. Contiene la decisión 347 que modifica la decisión 321, la decisión 351 sobre el régimen común sobre derecho de autor y derechos conexos, la decisión 352 que prorroga el plazo para poner en vigencia las modificaciones a la NANDINA, entre otras decisiones Lima, diciembre 21 de 1993.
- ✓ CONCEPTO DNDA. Concepto de la Dirección Nacional de Derecho de Autor (DNDA) sobre la presunción de transferencia del derecho patrimonial del autor (Derecho patrimonial de obra por encargo).

La base legal está determinada por los derechos de autor como un elemento trascendental cuando se trata de utilizar una fuente investigativa como trabajos de grado como base un corpus, dado que la legislación no brinda soluciones claras, algunos países han llegado a un consenso para otorgar ciertos privilegios a las universidades aun así la normatividad o limites sobre copia y uso de texto de corte investigativo universitario capturado a través de Internet tampoco está restringido, ni está restringido el número de palabras o líneas que se pueden copiar para no violar las regulaciones de derechos de autor.

Si bien es necesario y justo proteger los derechos de los autores y editores sobre los textos que crean o publican mediante derechos de autor es sumamente importante revisar y ampliar las regulaciones existentes en respuesta al rápido desarrollo de la tecnología de captura de texto computarizada. Puede ser probable que cualquier investigación que deba ser computarizado e

incluido en el corpus esté dentro del alcance de la ley y requiera autorización para su uso, y aunque la legislación varía de un país a otro como regla general, la duración de los derechos es limitada, y, en el caso de la cesión de los derechos, los propietarios de los derechos de autor tienen que tener la seguridad de que la compilación del corpus no será inconveniente para el potencial de ganancias y de que no habrá ninguna explotación comercial directa del corpus.

2.4 Definición de Términos Básicos o Glosario

A continuación, se presentan las palabras más relevantes contenidas en el desarrollo del trabajo de investigación ordenadas por orden alfabético.

- ✓ **Análisis de sentimientos**: Se refiere al uso de procesamiento de lenguaje natural, análisis de texto y lingüística computacional para identificar y extraer información subjetiva de unos recursos. (Liues, 2007)
- ✓ **Aprendizaje automático**: Es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. (Russell & Norvig, 2009)
- ✓ **Archivo/colección (informatizado) (*Archive/Collection*).-** Es un repertorio de textos en soporte informático sin buscar ningún tipo de relación entre ellos. (Percia, 2008)
- ✓ **Biblioteca de Textos Electrónicos (*Electronic text library*).-** Es una colección de textos en soporte informático, guardados en un formato estándar, siguiendo ciertas normas de contenido, pero sin un criterio riguroso de selección. (Percia, 2008)
- ✓ **Componente.-** Es una colección de muestras de un corpus o de un subcorpus, las cuales responden a un criterio lingüístico específico muy concreto. Los componentes reflejan un tipo determinado de lengua. Sobre todo, los corpus, pero también los subcorpus, son muy heterogéneos, mientras que los componentes son muy homogéneos. (Torruella & Llisterri, 1999)
- ✓ **Corpus lingüístico informático -** Es una recopilación de textos seleccionados según criterios lingüísticos, codificados de modo estándar y homogéneo, con la finalidad de poder ser tratados mediante procesos informáticos y destinados a reflejar el comportamiento de una o más lenguas codificado de manera estandarizada y homogénea para tareas de recuperación abiertas. (Torruella & Llisterri, 1999)

- ✓ ***Data mining***: Se refiere al proceso de descubrir patrones en grandes volúmenes de información, haciendo uso de inteligencia artificial, aprendizaje automático y estadística. (Maimon & Rokach, 2010)
- ✓ ***Dataset***: Hace referencia al conjunto de datos utilizado para entrenar un algoritmo de aprendizaje automático. (2021)
- ✓ ***Hardware y software***: Es la infraestructura informática de corpus, el componente de hardware (aparatos) y el de software (programas) son aspectos muy importantes al diseñar un corpus para poder desarrollarlo y explotarlo.
- ✓ ***Information retrieval***: Es una ciencia dedicada a la búsqueda de información en documentos electrónicos y cualquier tipo de colección documental digital. (Gerald & McGill, 1983)
- ✓ **Inteligencia artificial**: Es la inteligencia exhibida por máquinas. En ciencias de la computación, una máquina inteligente ideal es un agente racional flexible que percibe su entorno y lleva a cabo acciones que maximicen sus posibilidades de éxito en algún objetivo o tarea. (Poole, 2019)
- ✓ **Lingüística de corpus:** La lingüística de corpus es una rama de la lingüística que basa sus investigaciones en datos obtenidos a partir de corpus, esto es, muestras reales de uso de la lengua. En rigor, el término no define una disciplina lingüística, como lo pueden ser la morfología, la sintaxis o la pragmática, sino un enfoque metodológico que es posible adoptar desde disciplinas diversas, que se contrapone a una metodología basada fundamentalmente en la introspección. (Martín Peris, Cortés Moreno, Atienza, & López-Ferrero, 2008)
- ✓ **Lematización:** es un proceso lingüístico que consiste en, dada una forma flexionada como plural, femenino, conjugada, etc encontrar el lema proporcionado que es la forma que se acepta como representante de todas las formas flexionadas de una misma palabra, es decir que conduce a cada palabra a su raíz léxica ej. peligroso: peligro. (Fadić, 2020)
- ✓ **Lenguaje natural:** Se refiere a la lengua o idioma hablado o escrito por humanos para propósitos de comunicación. (Colmenarez, 2002)
- ✓ **Procesamiento de lenguaje natural PNL:** Es un campo de las ciencias de la computación, inteligencia artificial y lingüística que estudia las interacciones entre las computadoras y el lenguaje humano. El PLN se ocupa de la formulación e investigación de

mecanismos eficaces computacionalmente para la comunicación entre personas y máquinas por medio de lenguajes naturales. (Colmenarez, 2002)

- ✓ **Subcorpus.-** Suele ser una selección estática de textos, derivada de un corpus normalmente más general y complejo, el cual está dividido en grupos de muestras textuales más específicas; pero también puede ser una selección dinámica de textos de un corpus en crecimiento: un número determinado de textos destinados a aumentar algún apartado de un corpus general. (Torruella & Llisterri, 1999)
- ✓ **Sublenguaje**: Es un conjunto de piezas de lenguaje que se seleccionan y ordenan según un conjunto de criterios lingüísticos que sirven para caracterizar su homogeneidad lingüística. Mientras que un corpus puede ilustrar heterogeneidad. (Torruella & Llisterri, 1999)

CAPÍTULO III: ASPECTOS METODOLÓGICOS.

3.1 Paradigma

La investigación se enmarca en el paradigma cualitativo (Veron, E, 1987) dado el fin último que alude a las vinculaciones, en el significado del lenguaje verbal donde se interpreta las razones de los diferentes aspectos de un sentimiento, penetrando en el sentido profundo; en otras palabras, lo cualitativo interpreta e indaga en cuanto al por qué y el cómo; por lo cual se hace necesario además el complemento cuantitativo para la verificación del conocimiento objetivo , donde la racionalidad inherente está fundamentada en el cientificismo como postura epistemológica. (Hurtado & Toro, 2001)

3.2 Enfoque

El enfoque que se va a utilizar por las características presentes en el tema de investigación, es un enfoque hermenéutico utilizando la hermenéutica como abordaje cualitativo, en el cual se genera una nueva comprensión del saber cómo el método general de la comprensión e interpretación; el abordaje cualitativo hermenéutico es además multimetódico naturalista e interpretativo, con la indagación de situaciones naturales en el contexto social, a fin de interpretar los sentimientos en términos de los significados que las personas les otorgan, utiliza la neutralidad valorativa como criterio de objetividad. (Hernández R, et al., 1996) Aunado a este enfoque se aplicará el complemento cuantitativo el cual está sustentado en un enfoque descriptivo de la metodología que se usará en el proyecto de investigación.

3.3 Método

Se utilizará un método cualitativo hermenéutico, dado que, en el contexto de la investigación educativa, nos situamos en una de las perspectivas que distinguen Latorre, Del Rincón y Arnal (1996), “la perspectiva orientada a la práctica educativa, con una aplicación directa en la política o prácticas educativas y centrada en aportar información y herramientas que guíen la toma de decisiones y los procesos de cambio” (Hurtado & Toro, 2001). Para el presente proyecto estaríamos hablando de un tipo de investigación-acción, que sigue procesos de intervención en la práctica de investigación cualitativa y cuyo objetivo es transformarla. El estudio que aquí se presenta se enmarca también dentro de la investigación empírica, usando métodos cualitativos ya que pretende describir la aplicación de corpus lingüístico de la Maestría en Educación de la universidad de Nariño para la detección de sentimientos y categorías a través de la investigación

en el programa y descubrir sus ventajas, a la vez evaluar su funcionamiento. Para aquellos complementos cuantitativos, el método apropiado es experimental dado que se parte de una modelación para retomar resultados basados en aplicación directa en el fenómeno a estudiar.

3.4 Instrumentos Metodológicos

3.4.1 Fuentes Primarias.

Las fuentes de información primaria Trabajos de grado desarrollados dentro del marco de la Maestría de Educación de la Universidad de Nariño entre los años 2017 y 2021.

3.4.2 Fuentes Secundarias.

Como fuentes de información secundaria se considerarán la investigación en consultas, libros, revistas, tesis, encuesta sobre usabilidad del corpus y demás documentos que sirvan de apoyo al estudio propuesto.

3.4.3 Técnicas.

3.4.3.1 Ejercicio hermenéutico. dado qué, el propósito del estudio investigativo queda establecido a partir de la revisión bibliográfica de cada texto o tesis usada para el proceso de investigación, es necesario llegar hasta las profundidades del contenido, constantemente identificando los elementos que se van encontrando y dándole estructura, buscando especificar sentimientos, propiedades, características y rasgos importantes de cualquier fenómeno encontrado, describiendo las categorías, las subcategorías y tendencias respectivas que complementen de manera única el conglomerado del contenido encontrado para su estudio y comprender la esencia pura y vital de la hermenéutica, de captar el verdadero sentido mediante los diversos medios de verificación, tales como ver, leer, escuchar o sentir la verdad del emisor.

3.4.3.2 *Machine Learning*. La Lingüística de corpus genera una serie de métodos de investigación, tratando de trazar un camino de datos a la teoría, mediante el *Machine Learning* se permite entrenar modelos e inferir patrones del lenguaje mediante las 3ª; (Wallis, S. & Nelson G., 2001) introdujeron por primera vez lo que ellos llamaron la perspectiva de las tres A (3A *perspective*): anotación, abstracción y análisis para que los modelos de aprendizaje automático procesen el lenguaje con cada vez más precisión, entrenar modelos de aprendizaje supervisado y detectar determinados fenómenos lingüísticos.

CAPÍTULO IV: ANÁLISIS E INTERPRETACIÓN DE RESULTADOS

Para la construcción del corpus se debe seguir una metodología ordenada, mecánica y minuciosa. A continuación, se muestra paso a paso la metodología aplicada para lograr este objetivo:

4.1 Establecer Bases Conceptuales de Selección del Corpus o de los Textos que lo Van a Componer en el Plano Conceptual

Para llevar a cabo este desarrollo fue necesario primero establecer bases conceptuales o selección del corpus o de los textos que lo van a componer en el plano conceptual, a partir de estos se realizó un vaciado bibliográfico y lectura de las publicaciones del programa de maestría extraídos de la biblioteca de la Universidad de Nariño, con especial exhaustividad en la parte de aplicación al corpus.

Elegir las fuentes es una de las cosas importantes para tener en cuenta a la hora de comenzar la construcción de un corpus ya que deben ser fuentes fiables y de buena calidad lingüística para esto se realizó una minuciosa lectura de cada una de las tesis, con el fin de extraer frases que se acercan a la detección de sentimientos y categorización de los trabajos analizados. Las tesis que se tuvieron en cuenta se describen en la siguiente tabla:

Tabla 2 Listado de tesis de maestría para conformar el corpus lingüístico

#	Título	Año
1	LA FORMACIÓN EN VALORES ENTRE EL MITO Y LA REALIDAD	2017
2	REPRESENTACIONES SOCIALES DE LOS ESTUDIANTES ACERCA DEL MANEJO DE LA DIFERENCIA EN EL AULA	2017
3	COMPETENCIAS CIENTÍFICAS PROPICIADAS POR LA INVESTIGACIÓN COMO ESTRATEGIA PEDAGÓGICA (IEP) EN EL ÁREA DE CIENCIAS NATURALES Y EDUCACIÓN AMBIENTAL	2018
4	LA FORMACIÓN INICIAL DE EDUCADORES INVESTIGADORES Y LA PRÁCTICA PEDAGÓGICA INTEGRAL E INVESTIGATIVA EN LA FACULTAD DE EDUCACIÓN DE LA UNIVERSIDAD DE NARIÑO	2018
5	LA IDENTIDAD PROFESIONAL DOCENTE DESDE LAS CONCEPCIONES DE LICENCIADOS EN FORMACIÓN DE LAS ÁREAS DE ARTÍSTICA, CIENCIAS HUMANAS Y CIENCIAS EXACTAS Y NATURALES DE LA	2018

#	Título	Año
	UNIVERSIDAD DE NARIÑO	
6	RUTA PEDAGÓGICA PARA LA MEDIACIÓN INSTITUCIONAL ENTRE LOS ESTUDIANTES DEL PROGRAMA DE LICENCIATURA EN EDUCACIÓN FÍSICA DE LA INSTITUCIÓN UNIVERSITARIA CESMAG	2018
7	SENTIDO DE LA IDENTIDAD PROFESIONAL DOCENTE SEGÚN LAS CONCEPCIONES DE LOS LICENCIADOS EN FORMACIÓN ADSCRITOS A LOS DEPARTAMENTOS DE LINGÜÍSTICA E IDIOMAS Y ESTUDIOS PEDAGÓGICOS DE LA UNIVERSIDAD DE NARIÑO	2018
8	EL ROL DE LA UNIVERSIDAD EN EL TEJIDO DE LA PAZ	2019
9	IUIAY KAMSAYKUG TANTALLI TEJIDO DEL PENSAMIENTO UNA PROPUESTA INGA DE EDUCACIÓN PROPIA	2019
10	CARACTERÍSTICAS DE LA FORMACIÓN MUSICAL EN LA INSTITUCIÓN EDUCATIVA MUNICIPAL MERCEDARIO DE PASTO EN EL GRADO 6° DE BÁSICA SECUNDARIA	2020
11	CONTRIBUCIÓN DE LOS PROYECTOS AMBIENTALES ESCOLARES – PRAE, A LA GESTIÓN AMBIENTAL LOCAL, EN EL MUNICIPIO DE SAN JUAN DE PASTO, DEPARTAMENTO DE NARIÑO	2020
12	LA INTERDISCIPLINARIEDAD EN LA INNOVACIÓN EDUCATIVA	2020
13	LAS CREENCIAS RELIGIOSAS Y LOS IMAGINARIOS CULTURALES DE LOS ESTUDIANTES DE EDUCACIÓN MEDIA DE EL MUNICIPIO DE EL TAMBO, Y SU INCIDENCIA EN LOS PROCESOS CREATIVOS ARTÍSTICOS	2020
14	UNA MIRADA AL APRENDIZAJE DEL INGLÉS DESDE LOS IMAGINARIOS SOCIALES	2020
15	CONCEPCIONES SOBRE DIVERSIDAD CULTURAL Y SUS IMPLICACIONES EN LA SITUACIÓN ACTUAL DEL AWAPIT: EL CASO DE LAS COMUNIDADES EDUCATIVAS QUE PERTENECEN A LAS INSTITUCIONES INDÍGENA TÉCNICA AGROAMBIENTAL BILINGÜE AWÁ Y OSPINA PEREZ	2021
16	LA COMPETENCIA ARGUMENTATIVA ESCRITA DESDE LA	2021

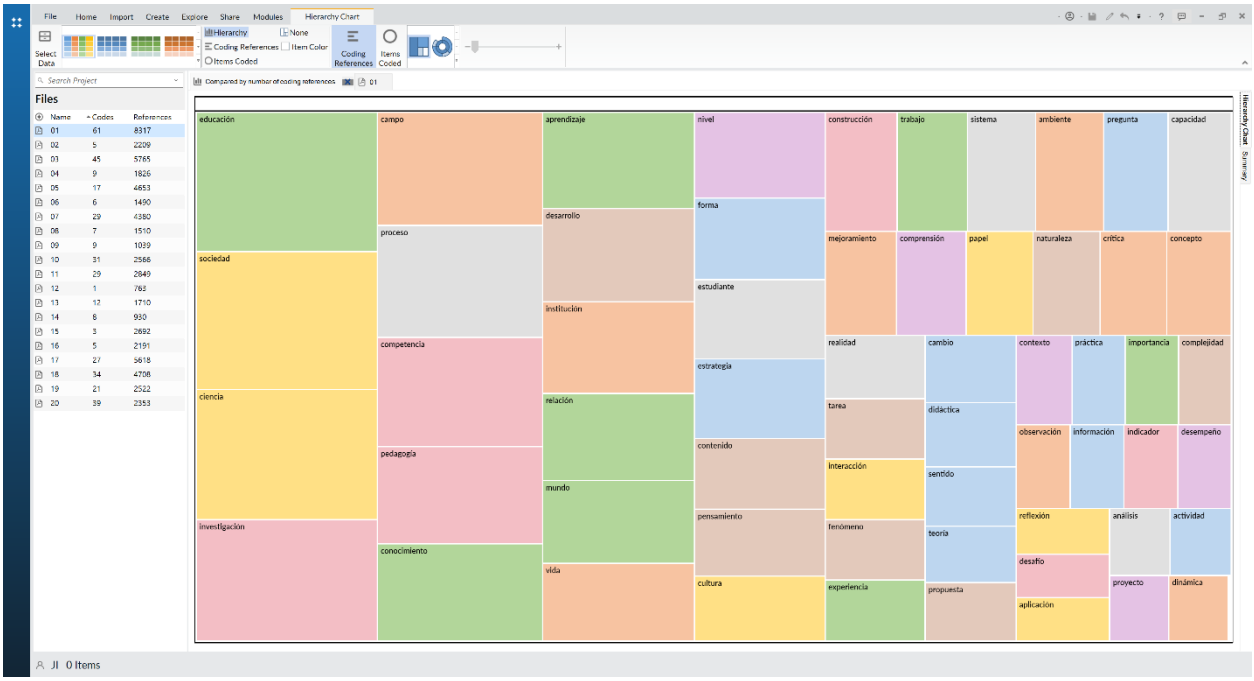
#	Título	Año
	PERSPECTIVA INTERDISCIPLINAR EN LOS ESTUDIANTES DE LA INSTITUCIÓN EDUCATIVA NORMAL SUPERIOR SAGRADO CORAZÓN DE JESÚS, SAN PABLO NARIÑO	
17	LAS PAUTAS DE CRIANZA Y LA VIOLENCIA ESCOLAR: UN ESTUDIO EXPLICATIVO EN UNA MUESTRA DE ADOLESCENTES ESCOLARIZADOS DE LOS MUNICIPIOS DE CHACHAGÜI Y PASTO	2021
18	PAUTAS DE CRIANZA REPRODUCIDAS POR LAS NIÑAS Y LOS NIÑOS DE PRIMER GRADO DE BÁSICA PRIMARIA FRENTE A LA CONCEPCIÓN DEL OTRO EN LA CONVIVENCIA ESCOLAR	2021
19	REPRESENTACIONES SOCIALES DE DOCENTES EN FORMACIÓN SOBRE DIVERSIDAD SEXUAL	2021
20	RETROALIMENTACIÓN DIVERSA PARA EL APRENDIZAJE DE LOS MODELOS ATÓMICOS	2021

Fuente: esta investigación.

4.2 Desarrollar una codificación y programación del material de entrenamiento de análisis lingüístico a partir del ejercicio hermenéutico de una muestra de las tesis de maestría en educación hasta la actualidad.

En este punto, se extrae información de una muestra de las tesis de Maestría en Educación de la Universidad de Nariño. Para ello se han tomado un compendio de 20 tesis de dicho programa y cada una de ellas fue sometida al procesamiento de información cualitativa por medios computacionales. Para tal efecto, se utilizó el software NVivo® de QSR International y de esta manera se apoyó el análisis hermenéutico sobre dichos documentos. Al introducir todas estas tesis que conforman el Corpus lingüístico del proyecto de NVivo, se procede A utilizar la función de auto codificación propia de la herramienta. Así cada tesis auto codificada, entendiéndose por codificación aquellas anotaciones a la margen de los documentos que hacen referencia a ideas principales detectadas por la herramienta a la luz de cada documento. la figura x muestra el resultado del proceso de auto codificación de una tesis.

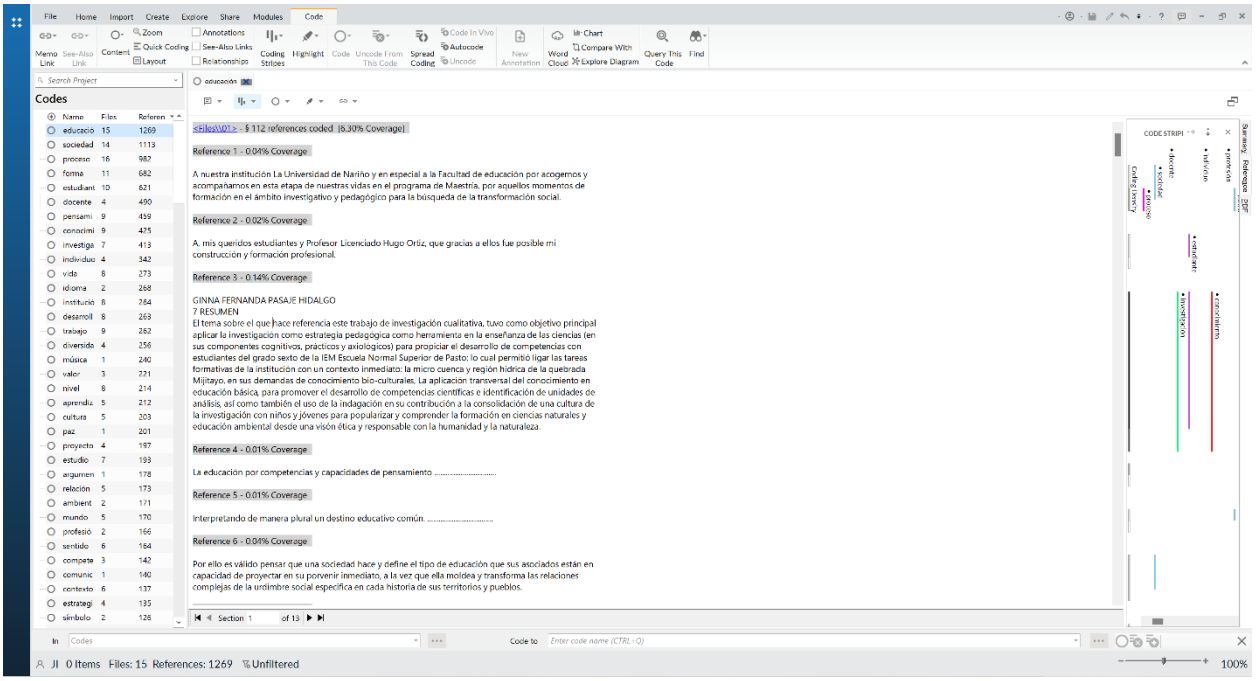
Imagen 2 Resultado de la auto codificación de una tesis



Fuente: esta investigación

Este proceso se repite por cada una de las tesis, así se obtiene un listado general de temáticas e ideas principales que gobiernan el contenido de cada documento que representa cada tesis de maestría. En esta parte se puede notar que cada auto codificación está asociada a fragmentos de texto al interior de cada tesis de maestría. Cuando se termina de hacer las auto codificaciones de todas las tesis de maestría si obtiene un listado de códigos e ideas comunes. al afinar y depurar dicho listado se procede a unificar las auto codificaciones comunes que han sido generadas a partir de todas las tesis. Resultado de esta unificación puede ser observado a través de la imagen 3.

Imagen 3 Unificación de auto codificaciones al texto de las tesis de maestría.



Fuente: Esta investigación.

Una vez obtenida la lista de códigos e ideas comunes que fueron extraídas a partir de las tesis de maestría se puede obtener una lista extensa de oraciones que semánticamente han sido asociadas a dicha codificación. la lista de códigos se muestra en la tabla 3.

Tabla 3 Lista de códigos e ideas comunes en las tesis de maestría

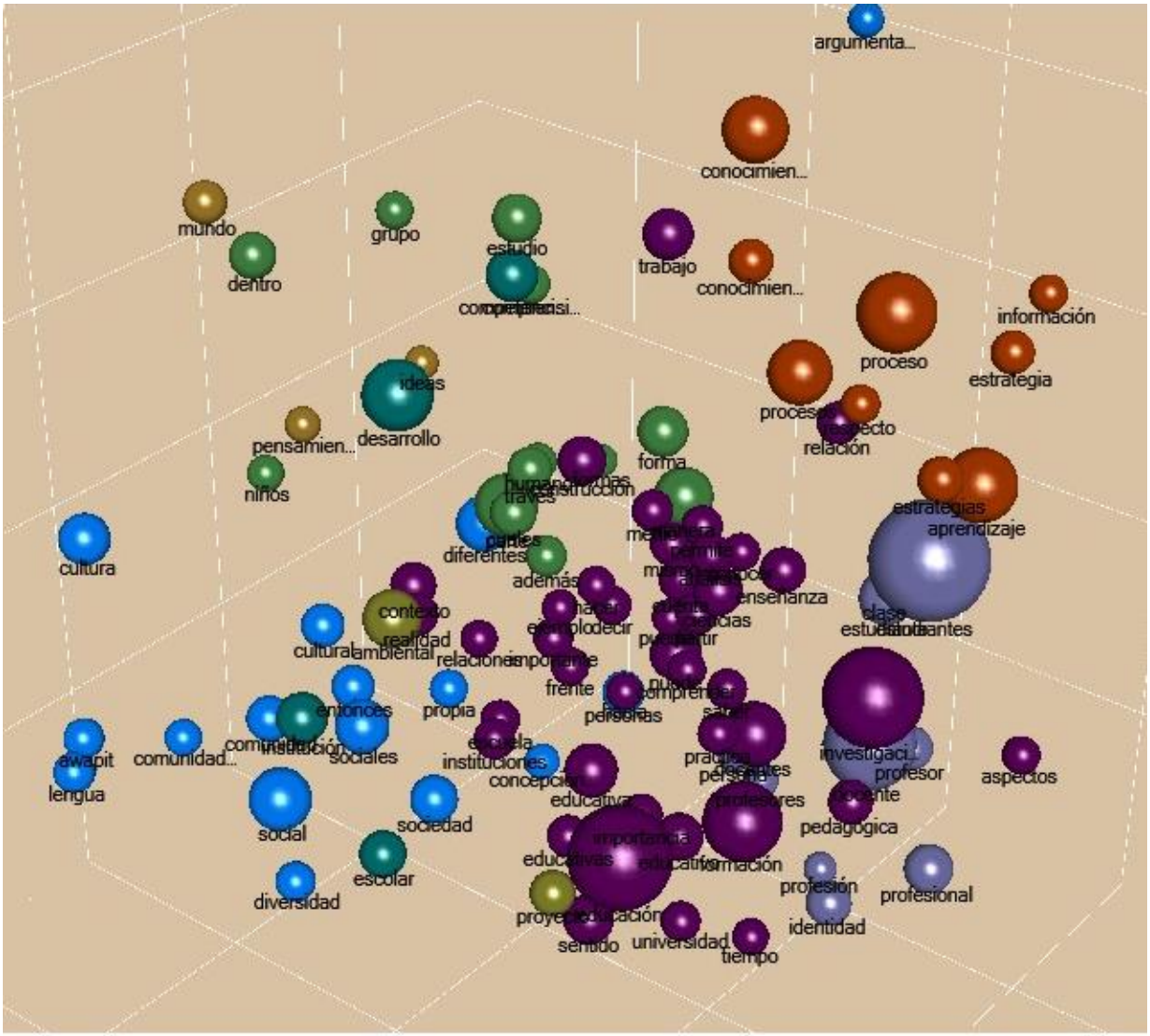
ACTIVIDAD	DERECHO	INTERACCION	PROPUESTA
AMBIENTE	DESAFIO	INTERPRETACION	PROYECTO
ANALISIS	DESARROLLO	INVESTIGACION	QUIMICA
APLICACION	DIDACTICA	LUGAR	REALIDAD
APRENDIZAJE	DINAMICA	MATEMATICA	REFLEXION
AREA	DIVERSIDAD	MEDIACION	RELACION
ARGUMENTACION	DOCENTE	MEJORAMIENTO	REPRESENTACION
ASPECTO	EDUCACION	METODOLOGIA	RESPUESTA
AULA	ELEMENTO	MODELO	RETROALIMENTACION
CAMBIO	ESPACIO	MUNDO	SENTIDO

CAMPO	ESPIRITUALIDAD	MUSICA	SEXO
CAPACIDAD	ESTRATEGIA	NATURALEZA	SISTEMA
CIENCIA	ESTRUCTURA	NIVEL	SOCIEDAD
CLASE	ESTUDIANTE	OBJETIVO	SOLUCION
COMPETENCIA	ESTUDIO	OBSERVACION	SÍMBOLO
COMPLEJIDAD	EXPERIENCIA	OPORTUNIDAD	TAREA
COMUNICACION	EXPRESION	PAPEL	TEMPLO
CONCEPTO	EVALUACIÓN	PAZ	TEORIA
CONDUCTA	FENÓMENO	PEDAGOGIA	TERRITORIO
CONFLICTO	GRADO	PENSAMIENTO	TEXTO
CONOCIMIENTO	IDIOMA	PERSPECTIVA	TIEMPO
CONSTRUCCIÓN	IMPORTANCIA	PRACTICA	TRABAJO
CONTENIDO	INDICADOR	PREGUNTA	TRATO
CONTEXTO	INDIVIDUO	PROBLEMA	USO
CREATIVIDAD	INFLUENCIA	PROCESO	VALOR
CRITICA	INFORMACION	PRODUCCION	VENTANA
CUALIDAD	INSTITUCION	PROFESION	VIDA
CULTURA	INSTRUMENTO	PROGRAMA	VIOLENCIA

Fuente: esta investigación

Esta lista de códigos e ideas comunes que constituyen el eje del Corpus lingüístico creado a través de las tesis de maestría está ligado a relaciones semánticas con el contenido de los documentos. De esta manera, los códigos e ideas comunes pueden ser representados en un escenario tridimensional a manera de *clusters* o conglomerados de acuerdo con sus relaciones semánticas al interior del Corpus lingüístico. dicha representación de conglomerados se puede observar a través de la imagen 4. En dicha figura se puede observar cómo los códigos e ideas comunes son representados a través de “atracciones gravitacionales” según su semántica. esto quiere decir que los conceptos se atraen cuando sus relaciones semánticas son más fuertes; esta es la esencia de los *clusters* o conglomerados.

Imagen 4 Representación tridimensional del conglomerado semántico de códigos e ideas comunes del corpus lingüístico



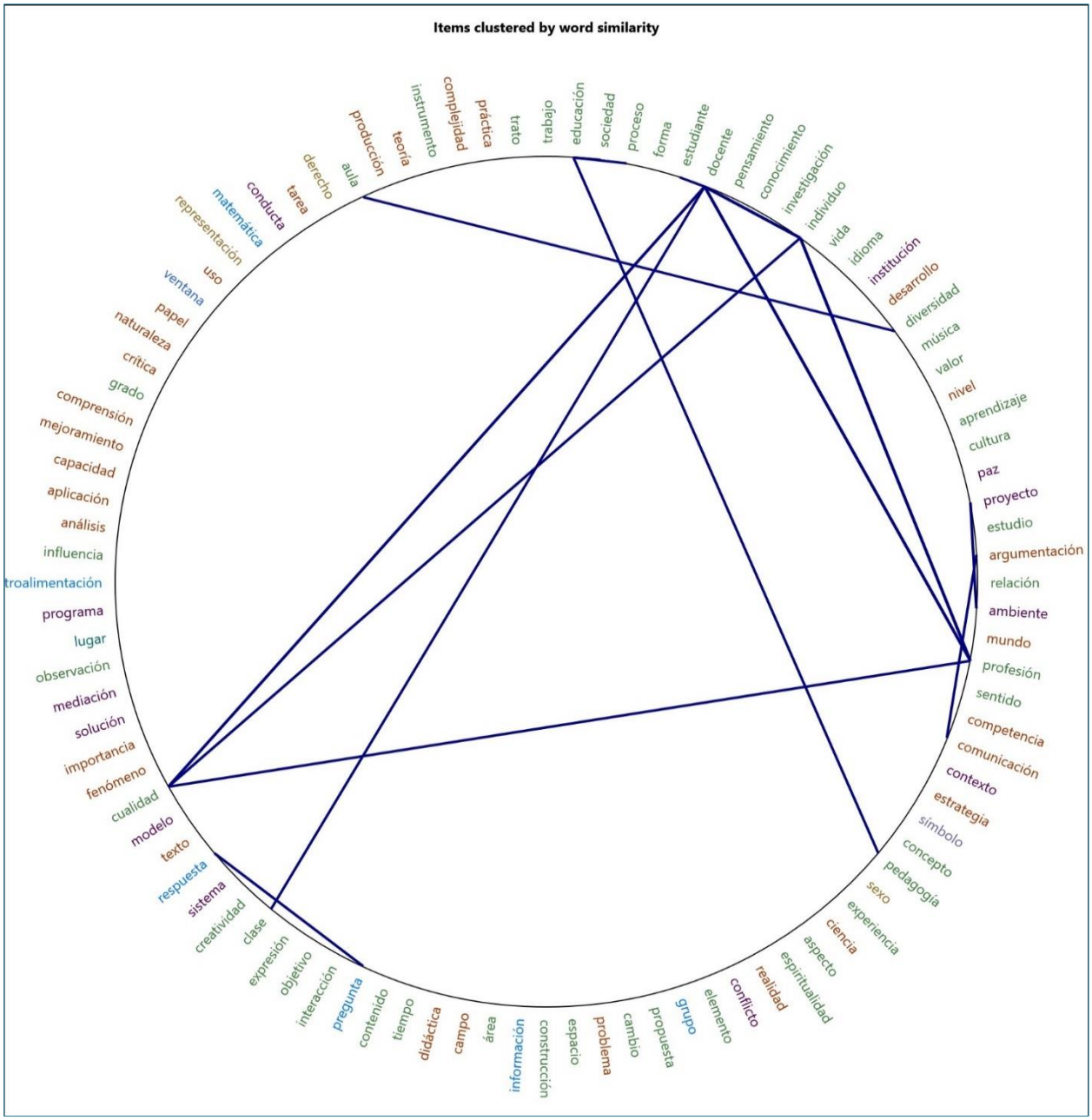
Fuente: Esta investigación.

Otra forma interesante de observar las relaciones de peso existentes entre los códigos e ideas comunes que han sido detectadas a partir del Corpus lingüístico es a través de una exposición circular de dicha codificación al tenor de su acercamiento semántico. en este orden de ideas la imagen 5 que muestra otra forma de representar dichas relaciones.

En esta disposición gráfica los colores década código e idea principal permiten agrupar semánticamente su cercanía. dentro de la disposición circular también existen Aces o rayos que

interconectan los conceptos asociados a los códigos e ideas principales que tienen gran fuerza y relevancia dentro del Corpus lingüístico.

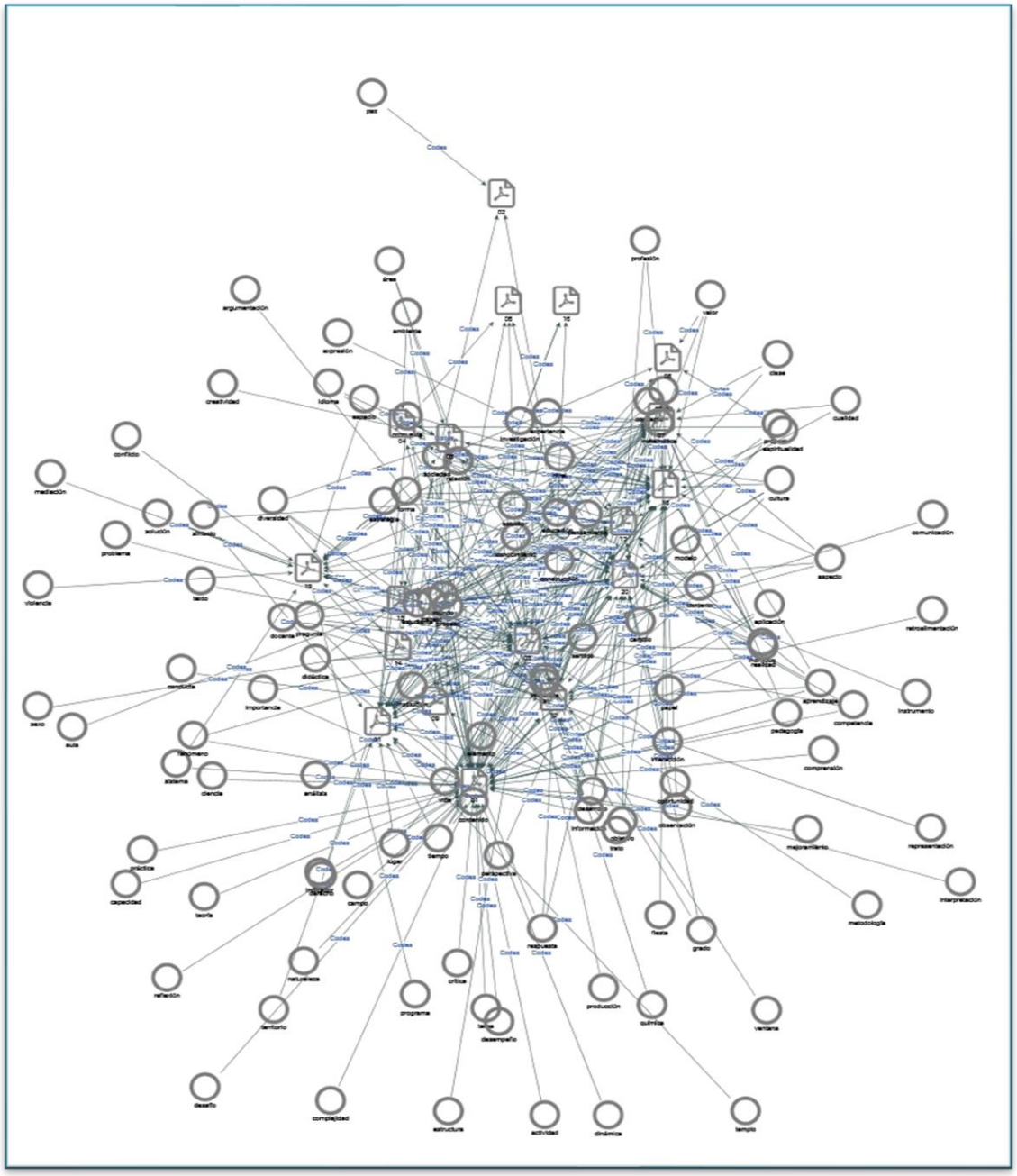
Imagen 5 Representación circular de las relaciones semánticas de los códigos e ideas principales del corpus lingüístico



Fuente: Esta investigación.

Finalmente, también es posible representar las relaciones semánticas entre los códigos y las ideas principales detectadas a partir del Corpus lingüístico a través de una red, para ello la imagen 6 presenta la información de cómo dicha relación puede ser representada a manera de una red de conexiones detalladas.

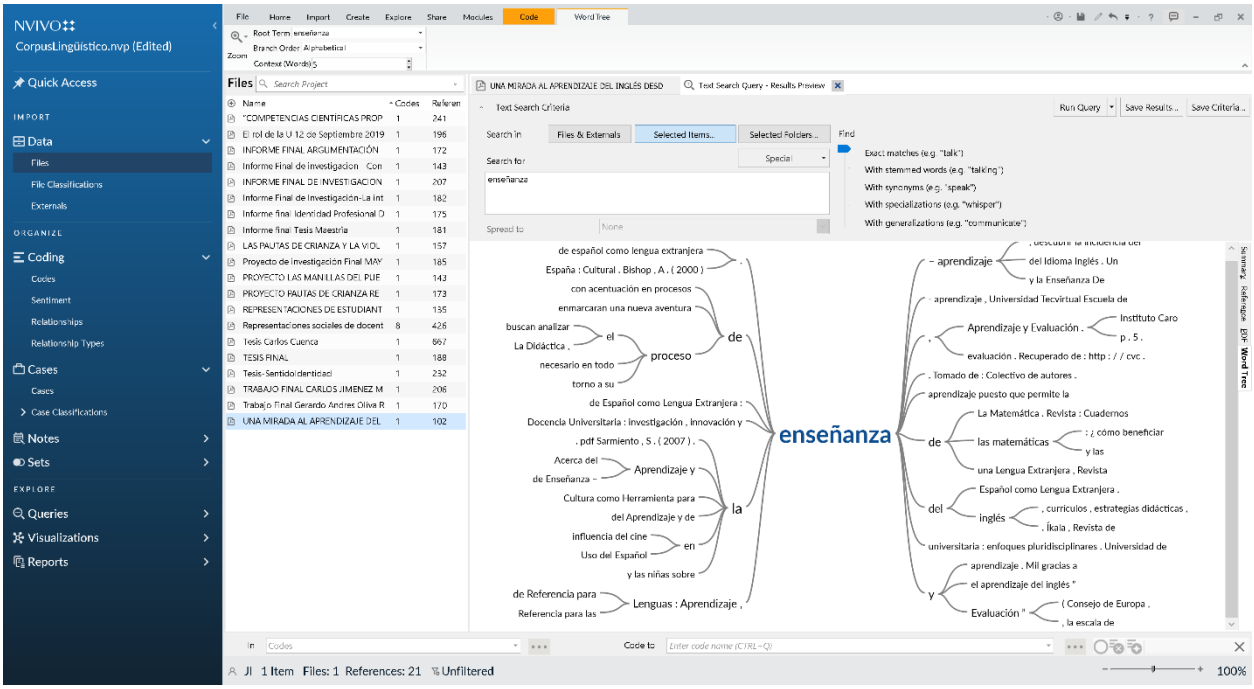
Imagen 6 Red semántica a partir de los códigos e ideas principales del corpus lingüístico.



Fuente: esta investigación.

En esta gráfica no solamente se integran las relaciones de los códigos e ideas principales, sino que también son relacionados los documentos PDF que en esencia son las tesis de maestría. Esta última representación ya tiene una mejor cercanía a lo que sería una propuesta de red neuronal con perceptrones basados en los códigos e ideas principales. esta última asociación se realiza internamente en la implementación de software desarrollada a través del grupo de investigación Galeras.net. Para tal efecto, el grupo de investigación hace uso de asistentes de *Machine Learning* que se despliegan en soluciones automatizadas de proyectos de esta naturaleza al interior de una herramienta denominada Microsoft Visual Studio. Para tal menester, las soluciones computacionales que son formuladas en Microsoft visual Studio consumen el componente de *machine learning* de .net denominado ML.net, Cuyo algoritmo de regresión logística se menciona posteriormente en este capítulo. Resultado de este procesamiento se obtiene una lista extensa de oraciones—aproximadamente 1,600 estructuras— que conforman los árboles gramaticales que representan la definición de sentimientos y categorías al tenor de cada investigación en dichas tesis. Esta forma de extraer las estructuras a partir de las oraciones seleccionadas se representa a través de la imagen 7.

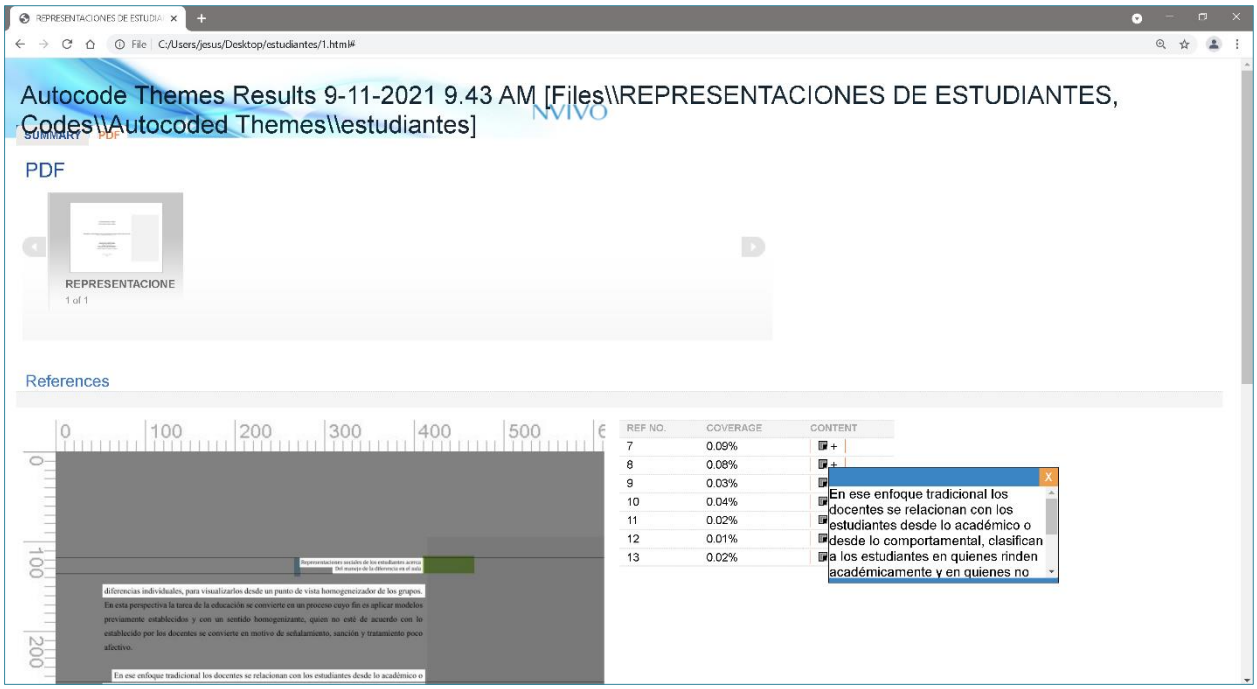
Imagen 7 Ejemplo de estructura gramatical extraída a partir de oraciones seleccionadas del corpus.



Fuente: Esta investigación.

El compendio de estructuras identificadas a partir de las oraciones seleccionadas se asocia a conceptos a través de análisis semántico. Producto de esta organización, se obtiene una lista semántica de oraciones con sus respectivas valoraciones como mecanismo de entrenamiento de la red neuronal programada por el Grupo de Investigación Galeras.Net del Departamento de Sistemas de la Universidad de Nariño. Dicha red neuronal está desplegada en un servidor de grupo y cuya configuración ha sido realizada para comportarse como un nodo computacional para *Machine Learning*. Un extracto de la lista semántica de oraciones con sus valores de entrenamiento para la red neuronal puede observarse en la imagen 8.

Imagen 8. Extracto de lista semántica de entrenamiento de la red neuronal.



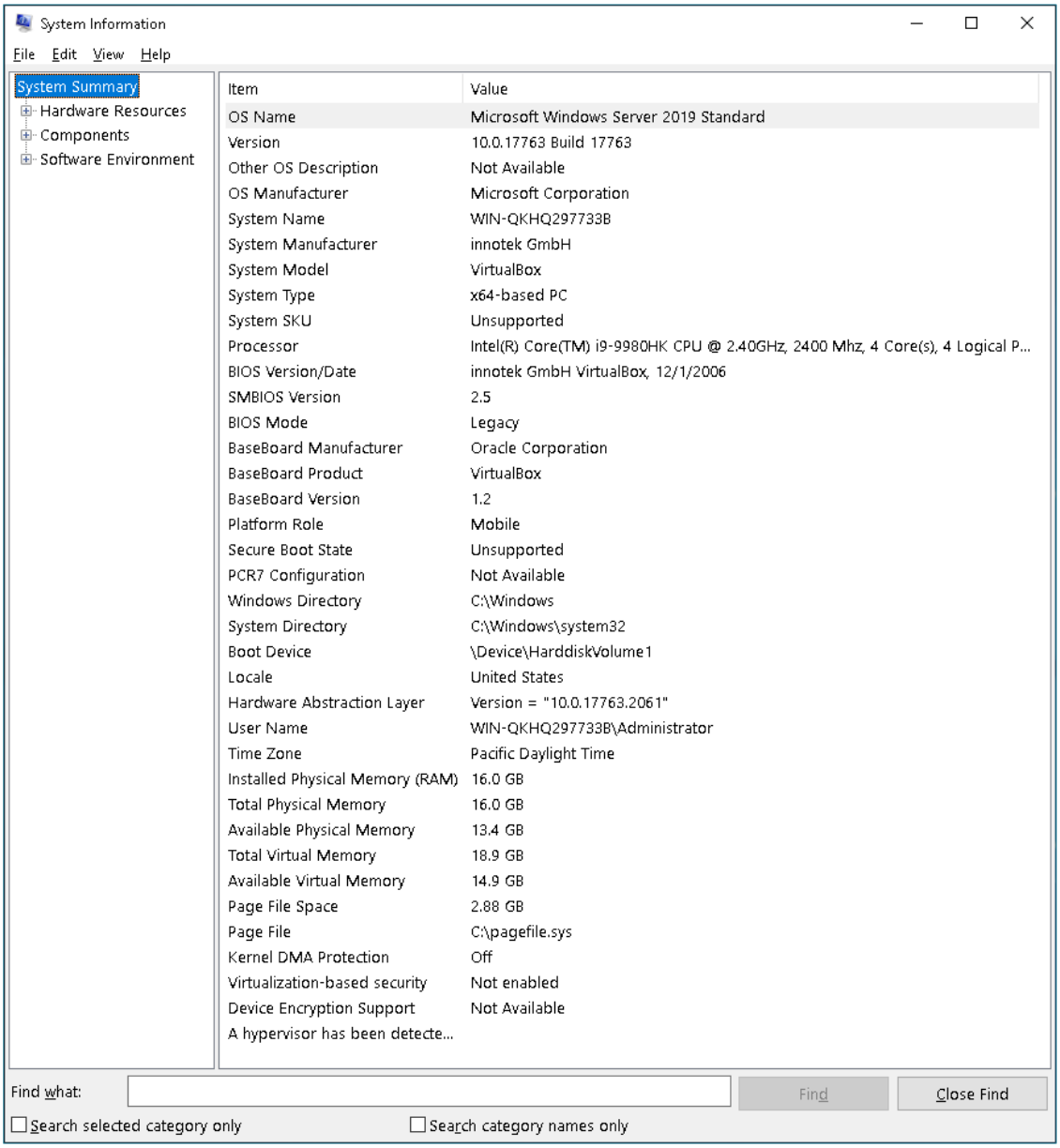
Fuente: Esta investigación.

4.3 Realizar el procesamiento informático de los textos para la materialización e implementación del soporte lógico y físico (*hardware y software*) del corpus.

Toda esta información que ha sido procesada a través de la herramienta de lingüística computacional NVivo fue exportada a través de archivos de Microsoft Excel los cuales constituyen los insumos de entrenamiento para el algoritmo seleccionado en la red neuronal. Aquí, se hace uso de la infraestructura del Grupo de Investigación Galeras.NET del Departamento de Sistemas de la Universidad de Nariño en el sentido de aprovechar los recursos

computacionales tales como servidores y software. Este conjunto de hardware y software constituyen la base fundamental de la solución computacional planteada. Desde el punto de vista del hardware, esta solución computacional cuenta con un servidor con las siguientes características descritas en la imagen 9:

Imagen 9. Características del servidor de despliegue de la solución computacional



Fuente: Esta investigación.

Las características de hardware anterior garantizan la ejecución de la solución computacional a través de técnicas de *Machine Learning*. Y es importante resaltar que los procesos de Machine learning pueden ser llevados a cabo en diferentes máquinas independientemente del hardware que tenga configurado, el sistema operativo a usar y el software que se pretenda utilizar.En este orden de ideas, el desarrollo de aplicaciones basadas en Machine learning no es exclusivo de un determinado hardware o de un determinado sistema operativo. el criterio utilizado por el grupo de investigación Galeras.net para usar este tipo de tecnología está basado en los recursos físicos que se tienen en el momento al disponer de licenciamiento académico basado en tecnología Microsoft, el grupo de investigación Galeras.net opta por utilizar el sistema operativo Microsoft Windows Server, El sistema de publicación WWW través de , y el sistema de desarrollo De Microsoft visual Studio por compatibilidad en este grupo de tecnologías.Una vez creado y desplegado el sistema de red neuronal a través del Grupo de Investigación Galeras.NET, se procede a insertar la lista de entrenamiento a dicha red para que a través de iteraciones de aprendizaje se puedan establecer los enlaces necesarios para la predicción de categorías y de sentimientos. Existe unos criterios generales acerca del tiempo de entrenamiento de una red neuronal dependiendo del tamaño del corpus lingüístico. En el caso particular de esta investigación, se ha logrado consolidar el tamaño del corpus lingüístico cerca de los 50 Mb de información. Por consiguiente, el tiempo de entrenamiento de la red neuronal (10 minutos o 600 segundos) corresponde a las recomendaciones que se especifican en la imagen 10, la cual resume el tiempo promedio necesario para obtener un buen rendimiento para un conjunto de datos de ejemplo que está representado por el corpus lingüístico en una máquina local.

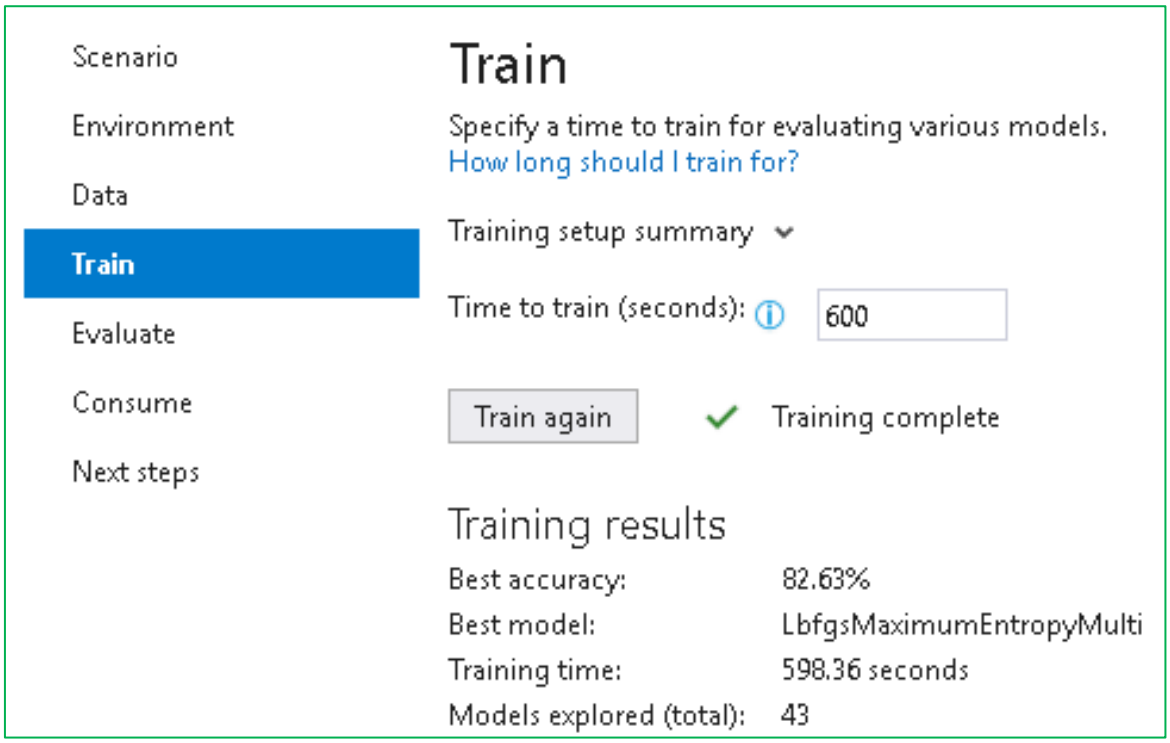
Imagen 10 Duración promedio del tiempo de entrenamiento de redes neuronales.

Dataset size	Average time to train
0 - 10 MB	10 sec
10 - 100 MB	10 min
100 - 500 MB	30 min
500 - 1 GB	60 min
1 GB+	3+ hours

Fuente: Adaptada de (Microsoft., 2020)

De acuerdo con la experiencia realizada, la red neuronal tuvo un entrenamiento de 10 minutos aproximadamente donde se pudo evaluar hasta 43 modelos basados en algoritmos preestablecidos. Esta información se puede observar a través de la imagen 11.

Imagen 11 Entrenamiento de la red neuronal a través de la lista de estructuras identificadas del corpus.



Fuente: Esta investigación.

La lista de resultados del entrenamiento se evidencia en la imagen 12.

Imagen 12 Experimentos de entrenamiento de la red neuronal y sus resultados

Experiment output folder: C:\Users\Administrator\AppData\Local\Temp\AutoML-NNI\Experiment-T2LXLN

	Trainer	MicroAccuracy	MacroAccuracy	Duration	#Iteration
0	SdcaMaximumEntropyMulti	0.5019	0.5000	4.0	0
1	FastForestOva	0.7396	0.7386	14.7	1
2	FastForestOva	0.7444	0.7434	14.8	2
3	FastTreeOva	0.6703	0.6690	8.9	3
4	LbfgsMaximumEntropyMulti	0.8088	0.8082	8.0	4
5	SdcaLogisticRegressionOva	0.5019	0.5000	5.9	5
6	LightGbmMulti	0.6724	0.6711	5.3	6
7	LbfgsMaximumEntropyMulti	0.8191	0.8182	4.6	7
8	SdcaMaximumEntropyMulti	0.5019	0.5000	3.0	8
9	LbfgsMaximumEntropyMulti	0.7838	0.7833	6.7	9
10	FastForestOva	0.7476	0.7463	11.4	10
11	LbfgsLogisticRegressionOva	0.8110	0.8104	7.8	11
12	FastTreeOva	0.6506	0.6491	7.9	12
13	LbfgsLogisticRegressionOva	0.7626	0.7626	6.8	13
14	LbfgsLogisticRegressionOva	0.8229	0.8220	7.5	14
15	LbfgsLogisticRegressionOva	0.7993	0.7990	6.7	15
16	LbfgsLogisticRegressionOva	0.8250	0.8239	7.6	16
17	LbfgsLogisticRegressionOva	0.8123	0.8116	22.7	17
18	LbfgsMaximumEntropyMulti	0.8216	0.8208	4.5	18
19	LbfgsMaximumEntropyMulti	0.8263	0.8251	10.1	19
20	LbfgsMaximumEntropyMulti	0.8119	0.8116	3.0	20
22	SdcaLogisticRegressionOva	0.5019	0.5000	6.6	22
23	SdcaMaximumEntropyMulti	0.5019	0.5000	3.2	23
24	LightGbmMulti	0.6851	0.6835	5.1	24
25	FastTreeOva	0.7214	0.7192	10.8	25
26	LightGbmMulti	0.4765	0.5000	4.8	26
27	SdcaMaximumEntropyMulti	0.5019	0.5000	3.1	27
28	FastTreeOva	0.7450	0.7464	10.9	28
29	SdcaMaximumEntropyMulti	0.5019	0.5000	2.9	29
30	LightGbmMulti	0.5924	0.6123	4.8	30
31	LightGbmMulti	0.6431	0.6511	4.4	31
32	FastTreeOva	0.6506	0.6491	9.7	32
33	SdcaLogisticRegressionOva	0.5019	0.5000	6.0	33
35	FastForestOva	0.7577	0.7579	15.2	35
36	SdcaMaximumEntropyMulti	0.5019	0.5000	2.9	36
37	LbfgsMaximumEntropyMulti	0.8080	0.8073	17.0	37
38	LightGbmMulti	0.7589	0.7579	5.9	38
39	LightGbmMulti	0.7095	0.7085	6.9	39
40	FastTreeOva	0.8052	0.8046	50.4	40
41	FastForestOva	0.7760	0.7756	62.6	41
42	LbfgsMaximumEntropyMulti	0.8177	0.8171	2.8	42

Fuente: Esta investigación.

Finalmente, el algoritmo para el modelo *FbfgsMaximumEntropyMultiClass* ha sido el mejor para la solución computacional planteada. Este algoritmo representa un modelo de entropía máxima el cual es una generalización de la regresión logística lineal. La principal diferencia entre el modelo de máxima entropía y la regresión logística es el número de clases admitidas en el problema de clasificación considerado. La regresión logística es solo para clasificación binaria, mientras que el modelo de entropía máxima maneja múltiples clases. (Yu, HF., Huang, FL. & Lin, CJ. , 2011)

4.4 Aplicar pruebas de validación de funcionamiento de corpus en la detección de sentimientos y categorías en los resultados de aplicación de instrumentos de recolección de información para las tesis de la maestría.

La solución computacional ha sido concebida como herramienta tecnológica para soportar los procesos de análisis hermenéutico en los resultados de aplicación de instrumentos de recolección de información en investigaciones cualitativas. Si bien es cierto que el Corpus lingüístico ha sido construido a partir de los documentos que conforman una muestra de las tesis de Maestría en Educación de la Universidad de Nariño, la solución computacional que representa la implementación de dicho Corpus lingüístico no necesariamente se restringe a dicho programa de posgrado.

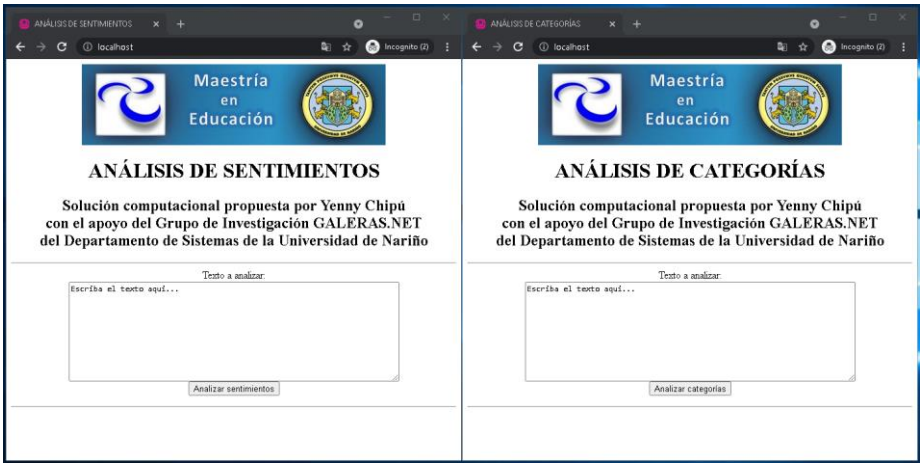
Por esta razón, y considerando las ventajas del uso de la tecnología en escenarios de investigación cualitativa, se han propuesto el diseño de interfaces gráficas de usuario preliminares a fin de dar funcionalidad a la solución computacional independientemente del escenario de aplicación. En este sentido, Lo que se trata es el aprovechamiento de la capacidad de cómputo que tiene la Universidad de Nariño, en este caso el grupo de investigación Galeras.Net, para que dichas soluciones sean desplegadas a nivel web y de esta manera poder facilitar un consumo masivo de los recursos.

A continuación, se presentan las interfaces gráficas que corresponden a páginas web que han sido publicadas usando los recursos del grupo de investigación. La imagen 8 permite evidenciar las interfaces gráficas para la detección de sentimientos y categorías respectivamente a través de entradas textuales que corresponden a los resultados de la aplicación de instrumentos de recolección de información de una investigación en curso.

Desde el punto de vista funcional, dichas interfaces son intuitivas ya que orientan al usuario a ingresar el texto a ser analizado en una caja de texto multilínea. Es importante considerar que entre más grande sea el texto, la red neuronal tardará más tiempo en procesar la información y realizar sus predicciones a través de su motor de inferencia.

A manera de ejemplo, se tiene una serie de textos que han sido resultado de la aplicación de instrumentos de recolección de información de algunas investigaciones que se encuentran en curso, algunos textos forman parte de una investigación en Maestría en educación y otros textos han sido tomados a partir de unas investigaciones en el programa de Maestría en Docencia Universitaria. La imagen 13 presenta esta situación de ingreso de valores.

Imagen 13 Interfaces Web de ingreso de texto para someterlo a análisis a partir de la lingüística de corpus.



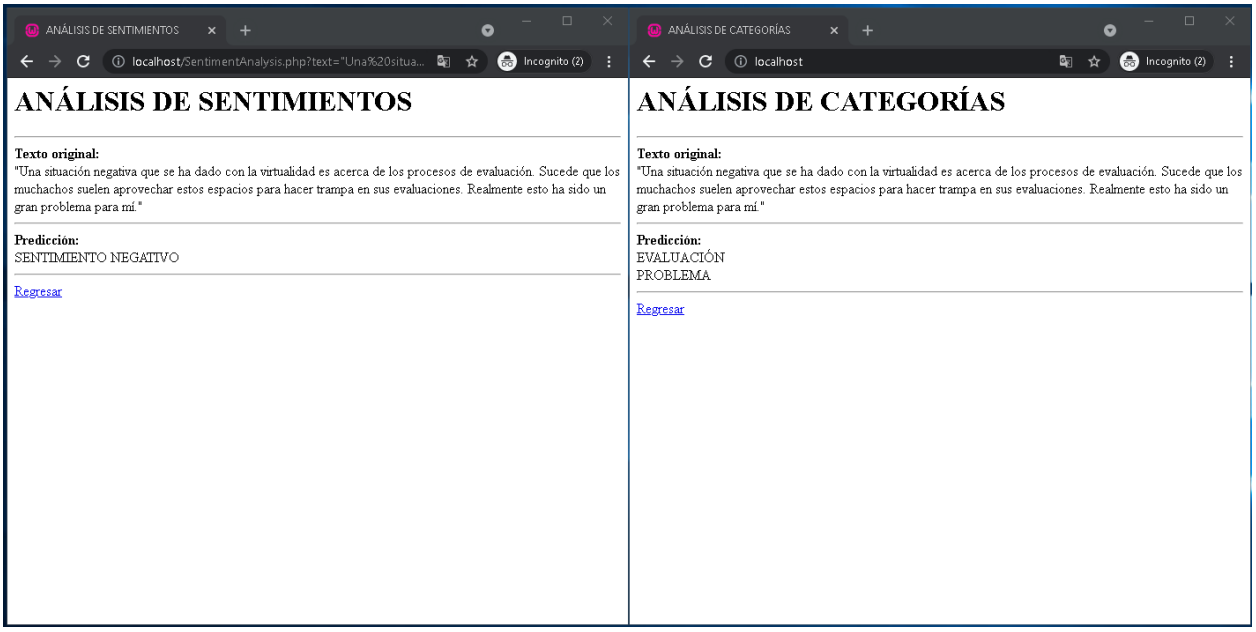
Fuente: Esta investigación.

Cuando se llama a las funciones de análisis, tanto para identificación de sentimientos como de categorías, la solución computacional devuelve los resultados tal como lo muestra la imagen 14. Las imágenes 14 y 15 representan los resultados de la funcionalidad del sistema en general. En el próximo capítulo se observará y el seguimiento realizado a una experiencia específica con 2 investigadores de diferentes programas los cuales sometieron textos de los resultados de la aplicación de los instrumentos de recolección de información de cada investigación, finalmente se analizará los resultados de dicha experiencia.

Imagen 14 Ingresando textos a ser analizados

Fuente: Esta investigación.

Imagen 15 Resultados del análisis lingüístico



Fuente: Esta investigación.

4.5 Análisis de los resultados de la investigación.

A continuación, se presenta un análisis de los datos obtenidos de las pruebas de validación de funcionamiento de corpus en la detección de sentimientos y categorías en los resultados de aplicación de instrumentos de recolección de información para las tesis de la Maestría.

4.5.1 Perspectiva del Producto

Considerando que la solución computacional está en un estado inicial, se tiene la posibilidad de extender su funcionalidad y su capacidad de acuerdo con los desarrollos que puedan realizarse a partir de esta solución construida. El hecho de que la solución computacional ha sido desplegada a nivel web, es una característica que puede hacer masivo su uso; La naturaleza de esta investigación apunta hacia el apoyo a los procesos de interpretación de los resultados de instrumentos de recolección de información que se puedan presentar en los escenarios de la investigación cualitativa. En este orden de ideas, la solución computacional puede ser consumida por diferentes tipos de usuario que tengan relación con la investigación cualitativa. Tan sólo en el contexto de la Universidad de Nariño se tiene una gran posibilidad de uso de la solución computacional más allá del programa de Maestría en Educación, ya que la Universidad tiene una oferta académica de programas tanto de pregrado como de posgrado que tengan relación con las

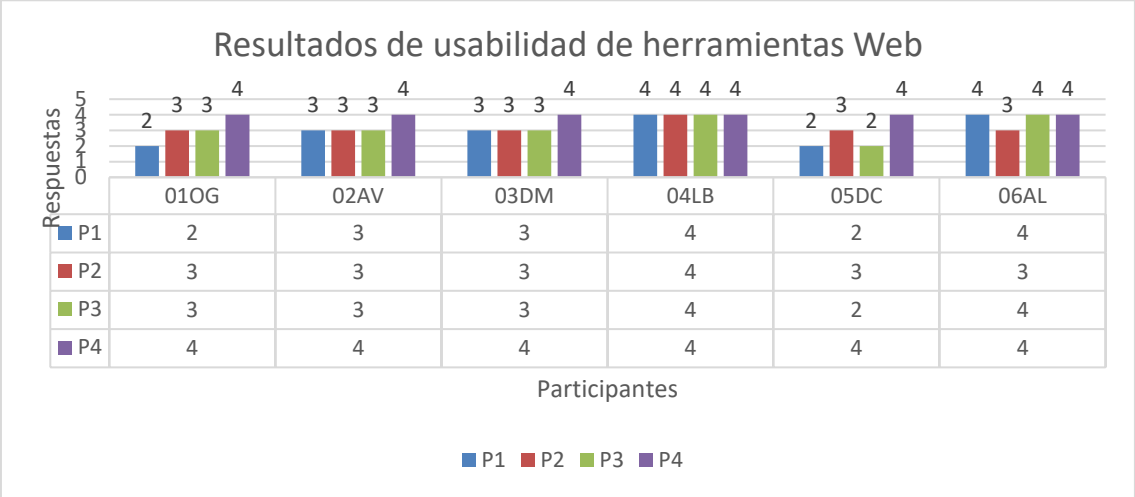
Ciencias Sociales, las Ciencias humanas, las artes y las letras, entre otros tipos de programas donde la investigación cualitativa se constituye como eje fundamental en consecuencia con la epistemología de dichas Ciencias. Así, las perspectivas del producto en cuanto a la posibilidad de uso y crecimiento de su funcionalidad es factible.

4.5.2 Funcionalidad del Producto

Las herramientas Web fueron probada a través de un estudio de usabilidad. Dicho estudio se basó en series de pruebas con diferentes grados de dificultad. 6 personas fueron requeridas; Según Nielsen, Turner y Lewis, (2006) "un pequeño número de usuarios encontrará la mayoría de los problemas" (p.3084-3088). Estas personas están desarrollando sus investigaciones cualitativas y disponen de algunos resultados de la aplicación de instrumentos de recolección de información. Cada participante realizó una prueba guiada de acuerdo con un proceso preestablecido.

Estas pruebas guiadas consistieron en 2 ejercicios de uso de las herramientas Web, una para la predicción de sentimientos y la otra para la predicción de categorías. Después de terminar cada prueba guiada, cada participante respondió una encuesta teniendo en cuenta la experiencia con las herramientas Web. Después de completar la prueba de usabilidad, los participantes llenaron una encuesta. Para la prueba de usabilidad, 6 estudiantes tuvieron interacción con las herramientas por un determinado tiempo de prueba entre los 15 y 20 minutos. Para ello, se pidió a los estudiantes que tomarán transcripciones de sus entrevistas y dichos fragmentos de texto fueron expuestos a las herramientas web. al procesar la información, las herramientas mostraban resultados sobre detección de sentimientos y sobre identificación de posibles categorías. los estudiantes repitieron este procedimiento hasta cumplir el tiempo pactado. al finalizar las experiencias, los estudiantes te respondieron a un instrumento de recolección de información.La gráfica 1 muestra el resumen de la Prueba de Usabilidad. Los resultados muestran que 6 personas participaron en esta prueba. Esta codificación se basa en números secuenciales del participante según el orden de la prueba (de 01 a 06) y las letras mayúsculas representan las iniciales de sus nombres. Los principales puntos planteados en la encuesta fueron:

Grafica 1 Resultados de la prueba de usabilidad en forma tabular



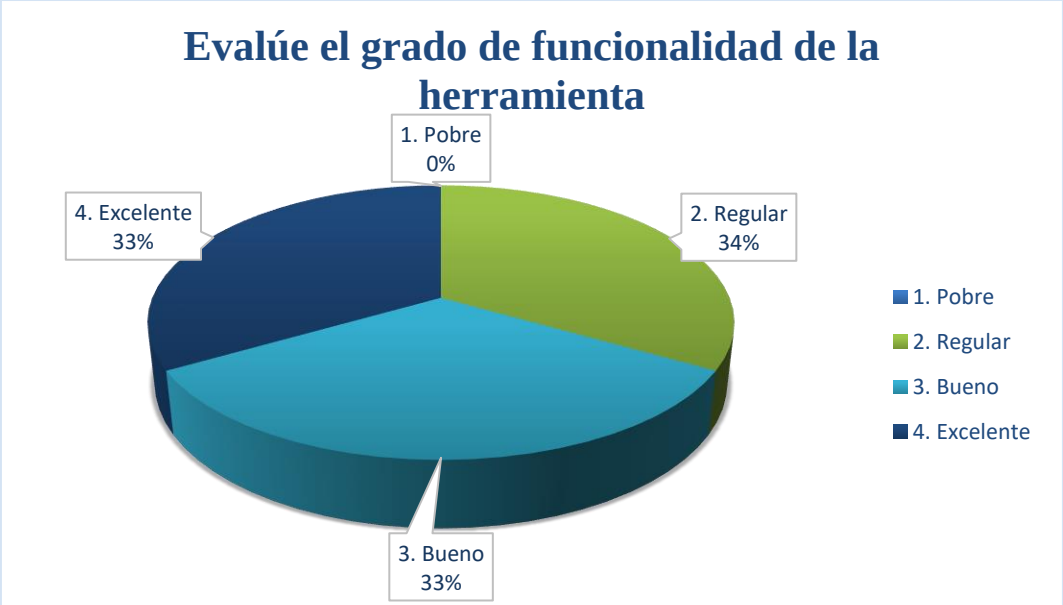
	01OG	02AV	03DM	04LB	05DC	06AL
P1	2	3	3	4	2	4
P2	3	3	3	4	3	3
P3	3	3	3	4	2	4
P4	4	4	4	4	4	4

Fuente: Esta investigación.

4.5.3.1 Tabulación y análisis de las encuestas

Una vez, se aplicaron las encuestas a los 6 estudiantes integrados de la siguiente manera: 3 de la XII promoción y 2 de la XIII promoción del programa de Maestría en Docencia Universitaria, y 1 del Programa de Maestría en Educación, se obtuvieron los siguientes resultados al realizar la respectiva tabulación, representación gráfica y análisis de las mismas:

Grafica 2 Pregunta 1. Evalúe el grado de funcionalidad de la herramienta



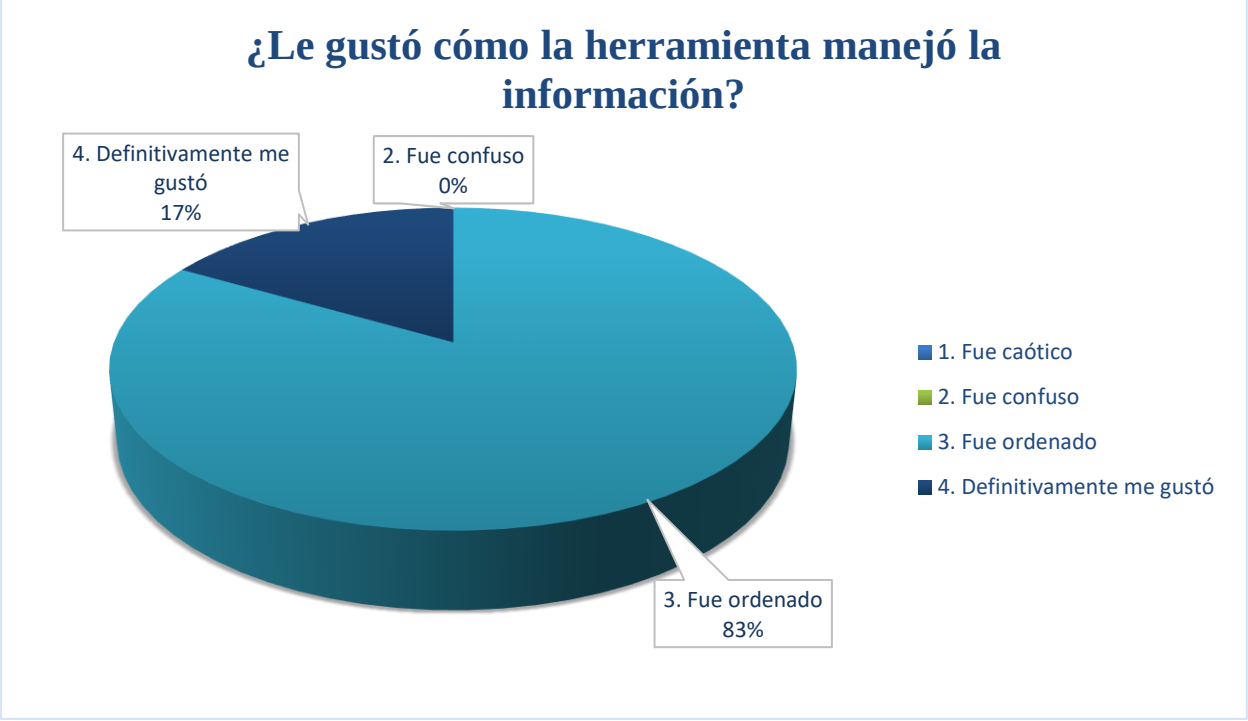
Fuente: Esta investigación. 2021

Respuesta	Número	Porcentaje
Pobre	0	0%
Regular	2	34%
Bueno	2	33%
Excelente	2	33%
Total	6	100%

Fuente: Esta investigación. 2021

La representación gráfica muestra que al preguntar sobre el grado de funcionalidad de la herramienta la evaluación por parte de los estudiantes es de 34% correspondiente a regular, y el 66% que corresponde a 4 de 6 estudiantes considera que el grado de funcionalidad de la herramienta es bueno y excelente.

Grafica 3 Pregunta 2. ¿Le gustó cómo la herramienta manejó la información?



Fuente: Esta investigación. 2021

Respuesta	Número	Porcentaje
Fue caótico	0	0%
Fue confuso	0	0%
Fue ordenado	5	83%
Definitivamente me gustó	1	17%
Total	6	100%

Fuente: Esta investigación. 2021

La gráfica muestra que frente a la pregunta sobre si le gustó cómo la herramienta manejó la información las respuestas obtenidas fueron de 83% que corresponde a 5 estudiantes afirman que en su opinión fue muy ordenado y 17% que corresponde a 1 estudiante afirmo que definitivamente le gusto como la herramienta maneja la información que se ingresa en ella.

Grafica 4 Pregunta 3. ¿Qué tan precisa es la herramienta?

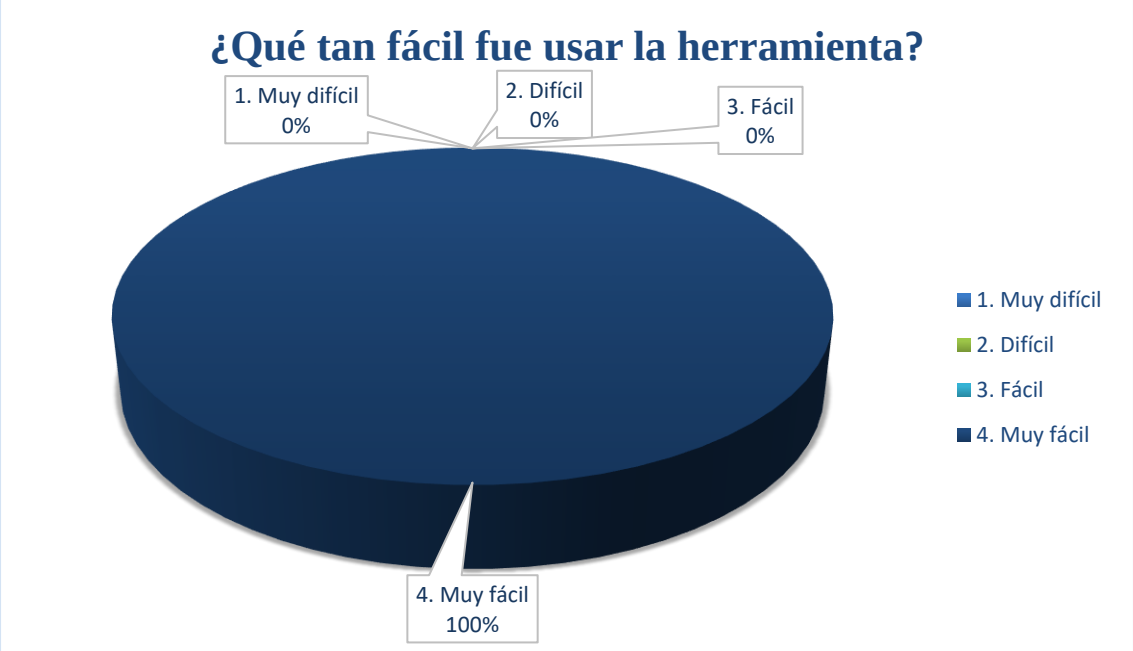
Fuente: Esta investigación. 2021

Respuesta	Número	Porcentaje
Muy imprecisa	0	0%
Imprecisa	1	17%
Precisa	3	50%
Muy precisa	2	33%
Total	6	100%

Fuente: Esta investigación. 2021

Al observar la gráfica, las respuestas a la pregunta sobre qué tan precisa es la herramienta las respuestas obtenidas son un 50% el cual corresponde a 3 estudiantes de 6 consideran que la herramienta es precisa y dos estudiantes que suman el 33% la consideran muy precisa, mientras que 1 estudiante que suma el 17% la consideró imprecisa.

Grafica 5 Pregunta 4. En general, ¿qué tan fácil fue usar la herramienta?



Fuente: Esta investigación. 2021

Respuesta	**Número**	**Porcentaje**
Muy difícil	0	0%
Difícil	0	0%
Fácil	0	0%
Muy fácil	6	100%
Total		100%

Fuente: Esta investigación. 2021

Teniendo en cuenta la gráfica anterior el 100% de los estudiantes que corresponde a 6 personas consideran que el manejo de la herramienta es muy fácil dado que su interfaz es muy intuitiva y accesible al usuario.

En general, la experiencia total en términos de usabilidad se consideró un buen resultado para las pruebas. Los hallazgos sugieren que las herramientas Web fueron bien aceptadas por los participantes. Teniendo en cuenta las respuestas, el común denominador otorgó un resultado positivo, lo que sugiere que las herramientas Web propuestas son factibles y altamente intuitivas.

4.5.3 Características de los Usuarios

Dado que el diseño de las interfaces gráficas para el ingreso de información textual hacer analizada a través de la solución computacional es intuitivo, lo que caracteriza a los usuarios de este tipo de sistema computacional es el análisis de información cualitativa de naturaleza textual. en este escenario no se requiere características especiales que deban disponer los usuarios para aprovechar las ventajas que ofrece la solución computacional. Esta solución computacional está orientada para quienes desarrollan investigación cualitativa en forma general, aunque el Corpus lingüístico ha sido conformado a partir de las tesis de Maestría en Educación de la Universidad de Nariño, esta característica propia de la solución computacional no restringe el tipo de usuario que interactúa con ella.

4.5.4 Restricciones

Teniendo en cuenta que la solución computacional planteada como apoyo a esta investigación ha sido diseñada y desarrollada con funcionalidades limitadas, el Corpus lingüístico resultado del

proceso de investigación tiene un tamaño pequeño. Esto hace que los análisis efectuados por la solución computacional tengan un cierto margen de error que se puede evidenciar en los resultados de la funcionalidad del producto. No obstante, es importante recordar que este tipo de soluciones computacionales que se apoyan en técnicas de *Machine Learning* no son 100% precisas; de hecho, todo sistema computacional es susceptible a la falla.

Una restricción importante que tiene la solución computacional es que su Corpus lingüístico ha sido conformado a partir de los textos escritos que conforman el informe final de investigaciones de tesis de maestría. En este orden de ideas, hay un uso de un lenguaje formal en dichos documentos. Al momento de realizar los análisis de sentimientos y categorías de los resultados de la aplicación de los instrumentos de recolección de información en algún tipo de investigación cualitativa, se observan ciertas falencias que corresponden al uso del lenguaje informal que se ha utilizado bien sea en encuestas entrevistas u otro tipo de técnicas de recolección de información.

4.5.5 Requerimientos Futuros

Considerando los resultados de la experiencia en escenarios reales, se hace necesario potenciar la solución computacional en el sentido de ampliar su capacidad de trabajo con texto plano; esto es, que la herramienta tenga la capacidad de procesar grandes volúmenes de información textual tal como los resultados compilados de todos los individuos quienes fueron sometidos a diligenciar instrumentos de recolección de información.

Por otra parte, se considera necesario incrementar la funcionalidad de la solución computacional en el sentido de permitir el análisis de sentimientos y categorías a fuentes no textuales de información. En este escenario, grandes ventajas podrían ser aprovechadas a partir del uso de la solución computacional al momento de realizar análisis lingüístico fuentes digitales multimedia de información tales como grabaciones en audio y vídeo de entrevistas, grabaciones de audio y vídeo de observaciones de campo como entre otros.

La solución computacional está disponible para el consumo masivo a través de sistemas web. Además, la solución computacional está alineada con los principios de publicación de software libre, lo que conlleva a la publicación de su código fuente abierto a fin de que pueda ser aprovechado por otros desarrolladores en el campo de la construcción del software. En este sentido, se tiene una gran expectativa frente a los desarrollos futuros que se puedan suscitar.

CONCLUSIONES

> *"La sociedad del conocimiento es también la sociedad del aprendizaje, que demanda aprendizaje a lo largo de toda la vida. En este contexto, las personas necesitan ser capaces de utilizar el conocimiento, de actualizarlo, de seleccionar lo que es apropiado para un contexto específico, de aprender permanentemente, y de entender el potencial de lo que aprenden, de tal forma que puedan adaptar el conocimiento a nuevas situaciones que se transforman rápidamente".* (Yániz, C. y Villardón Gallego, L. , 2006)

La presente investigación conduce a conclusiones de índole teórica y disciplinar y metodológico, puesto que hemos corroborado la utilidad, la pertinencia y las ventajas de la utilización de corpus lingüísticos a partir de tesis de la maestría de la universidad de Nariño, para ser más concretos, en la obtención de análisis de resultados enfocados y útiles sin sesgos de juicio para la investigación cualitativa, al mismo tiempo que enfatizamos la necesidad de disponer de las herramientas y soportes informáticos adecuados para que se puedan efectuar los distintos análisis de resultados en este tipo de investigaciones ya que el resultado de la unión entre la Lingüística Computacional y los corpus da lugar a avances científicos y tecnológicos en los que la gestión de la información y el tiempo han sido beneficiados, debido al desarrollo de la capacidad de almacenamiento, de análisis y de velocidad de respuesta.

El corpus provee la base para el análisis sistemático de una investigación cualitativas en términos de verificación objetiva de resultados, los beneficios de la utilización de *machine learning* para el análisis y la obtención de resultados están fundamentalmente avalados por los principios sobre los que se asienta la construcción del corpus lingüístico a partir de las tesis de maestría en Educación de la Universidad de Nariño para el apoyo hermenéutico en la detección de sentimientos y categorías en investigación cualitativa al ser una herramienta accesible, dinámica y sencilla de utilizar, además, los resultados obtenidos enriquecen las posibilidades de la metodología tradicional, atendiendo al volumen de los datos, tiempo de recopilación y procesamiento de la información.

Cañal (2005), señala: La presencia de las Nuevas Tecnologías en la sociedad y las potencialidades que éstas ofrecen como recursos para la educación e investigación constituyen una razón suficiente para justificar su incidencia en el perfil del profesor, en la medida en que

éste ha de desarrollar su acción educativa de un modo coherente con la sociedad en la que vive aprovechando al máximo los recursos que le ofrece. (Cañal, P., 2005) El docente universitario como formador de competencias en el saber y saber hacer, es tanto investigador como gestor tecnológico y líder, por lo cual debe conocer las nuevas tecnologías en todas sus dimensiones, lograr analizarlas de forma crítica y constructiva del saber investigador como recurso de la información, aunado a que la universidad tiene como propósito coordinar innovaciones educativas y tecnológicas para el uso del entorno virtual el corpus desarrollado tanto para investigadores, como para docentes y para estudiantes de las diferentes carreras, representa un potencializador de manejo de información y es un mecanismo ideal para cumplir con los niveles de excelencia educativa.

Por otra parte la virtualidad en el quehacer educativo se refiere a una dinámica de transformación constante que se da en un contexto con docentes y estudiantes orientados por la gestión académica hacia el objetivo del desarrollo formativo, las herramientas virtuales se convierten en la nueva normalidad del quehacer investigativo ajustada al entorno para la construcción de conocimiento, el desarrollo de esta herramienta permite acortar distancias en la labor de desarrollo de investigaciones cualitativas de forma interdisciplinaria, ya que, utiliza la red como canal de comunicación en los diferentes entornos tecnológicos como espacios cooperativos de investigación lo cual se resume, fundamentalmente, en la posibilidad qué la herramienta del corpus brinda al investigador obtener resultados intrínsecamente más verificables que los juicios basados en la introspección.

Desde el primer momento, se pretendía que la herramienta se caracterizara por la facilidad y la sencillez, accesibilidad rapidez, facilidad a la hora de su utilización por parte del investigador o cualquier tipo de usuario, desde el punto de vista informático, este aspecto quedó resuelto mediante el recurso de utilizar una herramienta metodológica de apoyo hermenéutico al servicio de la investigación cualitativa que puede ser adoptado desde diversas disciplinas mediante la web a la que se pudiera acceder desde cualquier ordenador con acceso a Internet.

Tanto en el sector académico como en el profesional, es de suma importancia no solo buscar maneras y recursos que permitan el desarrollo y construcción de corpus de calidad y el uso de software libre para manejarlos fácilmente, que permitan ampliar el conocimiento sobre su realización y uso en las diferentes disciplinas educativas.

RECOMENDACIONES.

Si bien el corpus construido nace a partir de las tesis de la maestría de Educación de la Universidad de Nariño, su utilización puede ser impulsada y globalizada a todo investigador, desarrollador o estudiante, así como también en el mundo profesional para los egresados de la Universidad de los diferentes programas.

Las líneas de trabajo futuro que derivan o están relacionadas directamente con el desarrollo del corpus recaen en la importancia que ha demostrado tener la colaboración entre grupos de investigadores y estudiantes de las diferentes carreras de la universidad de distintas áreas de conocimiento por lo cual es de suma importancia potenciar la creación de grupos de trabajo interdisciplinares que posibiliten el desarrollo de proyectos que sumen conocimientos de múltiples disciplinas para obtener resultados susceptibles de ser aplicados en los entornos más dispares, y no sólo en casos aislados o desarrollados por un grupo de trabajo de investigación específico.

Es vital impulsar a los estudiantes y a los docentes el uso de esta herramienta no solo como apoyo hermenéutico en las investigaciones de corte cualitativos sino además en el incentivar la posibilidad para desarrolladores o estudiantes para implementar funciones o complementación en formas de subir la información como de manera de audio, imagen, video o reconocimiento de voz grabada entre otras posibilidades por ser un corpus de software abierto.

BIBLIOGRAFÍA

McEnery , T., & Wilson, A. (1996). *lancaster.ac.uk.* (E. U. Press, Ed.) Retrieved Julio 8, 2021, from http://www.lancs.ac.uk/fss/courses/ling/corpus/

Russell, S. J., & Norvig, P. (2009). *Inteligencia Artificial: Un Enfoque Moderno* (3 edición ed., Vol. 2). (c. M. Universidad Pontificia de Salamanca, Ed.) PEARSON EDUCACIÓN, S.A.

Anguera, M. T. (1995). *"La investigación cualitativa".* Madrid, España: Educar.

Baeza, M. A. (2002). De las metodologías cualitativas en investigación científico social. Diseño y uso de instrumentos en la producción de sentido ". Concepción: Editorial de la Universidad de Concepción.

Beuchot, M. (2008). "Breve exposición de la hermenéutica analógica". *Revista Teología*(tomo XLV), 491-502.

Borreguero Zuloaga, M. y. (2021, 04 12). ¿Cómo hablan los jóvenes? Los corpus lingüísticos como base para la reflexión teórica y el aprendizaje de lenguas. Madrid, España: Universidad Computense de Madrid.

Campoy Aranda , T. J., & Gomes Araújo, E. (2009). *Técnicas e instrumentos cualitativos de recogida de datos.* Retrieved from http://proyectos.javerianacali.edu.co/cursos_virtuales/posgrado/maestria_asesoria_familiar/Investigacion%20I/Material/29_Campoy_T%C3%A9cnicas_e_instrum_cualita_recogidainformacion.pdf

Cañal, P. (2005). *La innovación educativa.* Andalucía, España: Akal. Retrieved Septiembre 1, 2021, from Disponible: http://ow.ly/bW5F304MmIi

Caravedo Barrios, R. (1999). *Apuntes metodológicos: Lingüística del corpus.* Salamanca: Ediciones Universidad de Salamanca.

CASACUBERTA, F. et al. (1992). *Grupo de Fonética del Departamento de Filología Española de la Universidad Autónoma de Barcelona.* (U. A. Barcelona, Ed.) Retrieved 08 15, 2021, from http://liceu.uab.es/~joaquim/publicacions/Casacuberta_et_al_92_Corpus_Albayzin.pdf

Chávez, B. L. (2014). Retrieved 06 12, 2021, from https://dx.doi.org/10.6018/analesps.30.3.154931

Colmenarez, J. (2002). *Introduccion a la computacion* . Universidad Fermin Toro.

Congreso de la República de Colombia. (1993, Febrero 5). LEY 44 DE 1993. Bogota, Colombia.

Congreso de la República de Colombia. (2019). Ley 1951 de 2019. Bogota, Colombia: Senado de la Republica.

Congreso de la República de Colombia. . (1982, 01). LEY 23 DE 1982. Sobre derechos de autor. . Bogota, Colombia.

Constitución política de Colombia. (1991). *Constitución política de Colombia Artículo 61 [Titulo II]*. (2da Edicion. ed.). Bogota, Colombia: Legis.

Cristian Giovanny Álvarez Calderón, F. A. (2019). El análisis lingüístico a través de un corpus de entrevista oral. *Trabajo de Fin de Grado*. Bucaramanga, Colombia: Universidad cooperativa de Colombia.

Dilthey, W. (1944). *El origen de la hermenéutica* (Vol. vol. V). México, México: Fondo de Cultura Económica - Obras de Wilhelm 172 Dilthey.

Fadić, M. N. (2020). *Corpus Básico del Español de Chile metodología de procesamiento y análisis.* (2 ed.). Santiago de Chile , Chile : Lexis 44.

Fernández, A. G. (2016). Más allá del corpus big data en la investigación lingüística. Evolución, analisis y predicción del uso de la lengua a través de twitte. Cordoba, España: Universidad de Córdoba (España).

FRANCIS, W. N. (1992). *Language Corpora B.C. en J. SVARTVIK* . Stockholm, Berlin/New York: Mouton de Gruyter.

Garrote Salazar , M. (2010). *Los corpus de habla infantil: metodología y análisis.* Madrid, España: Editorial Universidad Autónoma de Madrid. ProQuest Ebook Central.

Gerald, S., & McGill, M. (1983). *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.

Grondin, J. (2018). *“¿En qué consiste el sentido hermenéutico?” Entornos de la hermenéutica por los caminos de Jean Grondin.* (U. N. México, Ed.) México, México.

Hernández R, et al. (1996). *Metodología de la Investigación.* Ciudad México, México: Mc Graw Hill.

Hsinchun, C., & Chau, M. (2004). *“Web Mining: Machine Learning for Web Applications”* . . (B. Cronin, Ed.).

Hurtado, L., & Toro, G. (2001). *Paradigmas y Métodos de Investigación en Tiempos de Cambios. ;* (4ta ed ed.). (Epísteme, Ed.) Valencia, España: Consultores Asociados CA.

Juan Pedro Rica Peromingo, A. L. (2019, 03 30). *Corpus lingüístico y tecnologías para la enseñanza, el aprendizaje y la investigación en traducción audiovisual y accesibilidad*

lingüística (subtitulado para sordos, audio descripción para ciegos y Lengua de Signos Española): normativas de aplicación. Retrieved 06 12, 2021, from Universidad Computense de Madrid: https://eprints.ucm.es/id/eprint/57028/

Liues, B. (2007). *Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data.* Alemania, Colombia: Springer.

López Sanjúan , V. (2017, Abril 19). Integración de los corpus como herramienta de apoyo en la enseñanza de ESP. *Porta Linguarum*, 22. Retrieved Julio 10, 2021, from http://www.ugr.es/~portalin/articulos/PL_numero10/9%20Victoria%20Lopez.pdf

Maimon , O., & Rokach, L. (2010). *Manual de minería de datos y descubrimiento de conocimientos* (2a ed. ed.). (O. M. Rokach, Ed.) Editor Oded Maimon y Lior Rokach: Springer.

Martí Antonín , M., & Castellón Masalles, I. (2000). *Lingüística computacional,,.* Barcelona, España: Universitat de Barcelona.

Martín Peris, E., Cortés Moreno, M., Atienza, E., & López-Ferrero, C. (2008). *Diccionario de términos clave de ELE.*

McEnery, T. (2003). *"Corpus Linguistics", en R. Mitkov (ed.).* (O. U. Press, Ed.) Oxford: The Oxford Handbook of Computational Linguistics.

McEnery, T., Xiao, R. y Tono, Y. (2006). *Corpus-Based Language Studies. An advanced resource book, .* London/New York: Routledge.

Meyer, I. & Mackintosh, K. (1996). *The Corpus from a Terminographer's Viewpoint. En International Journal of* (Vol. vol. 1(2)).

Microsoft. (2020, Agosto 30). *'What is Model Builder and how does it work?'. .* Retrieved from Microsoft Documentation.: https://docs.microsoft.com/en-us/dotnet/machine-learning/automate-training-with-model-builder

Mikolov, T., Kai , C., Greg , C., & Jeffrey , D. (2013). *Regularidades lingüísticas en el espacio continuo Word representartaciones.* NAACL HLT .

Mitchell, T. (1997). *Machine Learning.* Nueva York: McGraw Hill.

Nielsen, J., Turner, C., & Lewis, J. (2006). *Determining Usability Test Sample Size.* (Vols. Vol. 12, No. 3.). International Journal of Ergonomics and Human Factors.

Nilsson, N. J. (1998). *Robotics Stanford.* Retrieved 10 25, 2021, from http://robotics.stanford.edu/people/nilsson/MLBOOK.pdf

Padilla, K. M. (2015). *«Los corpus lingüísticos al servicio de la semántica : su empleo en la delimitación de sentidos contextuales»*. (U. A. Barcelona, Ed.) Retrieved 06 12, 2021, from Dipòsit Digital de Documents de la UAB: https://ddd.uab.cat/record/174510

Percia, H. M. (2008). *Fundamentos de la lingüística de corpus* . Letras: Filología Hispánica.

Poole, D. (2019). *Computational Intelligence: A Logical Approach.* Nueva York: Oxford University Press.

Presidencia de la Republica de Colombia. (1991). DECRETO 393 DE 1991 . Bogota , Colombia .

Ricoeur, P. (1998). La teoría de la interpretación. Discurso y excedente de sentido.: . Madrid, España: Ed. Siglo XXI.

Rojo., G. (2002). *Centro de Terminología y Lexicografía.* (T. A. University-Texarkana, Ed.) Retrieved Julio 27, 2021, from www.uzei.com/Modulos/UsuariosFtp/Conexion/archivos54A.pdf

s.f. (2020). *Corpus lingüístico.* Retrieved 07 27, 2021, from Universidad Politecnica de Madrid: http://lorien.die.upm.es/juancho/pfcs/AJP/cap4.pdf

Santalla del Río, M. P. (2005). *“La elaboración de corpus lingüísticos”, en M. Cal, P. Núñez e I. M. Palacios (eds.): Nuevas tecnologías en Lingüística, Traducción y Enseñanza de lenguas* . (U. d. Compostela, Ed.) Servizo de Publicacións e Intercambio Científico.

Schleiermacher, F. D. (1829). *Sobre el concepto de hermenéutica.* Berlín, Boston: De Gruyter.

Sinclair, J. M. (1996). *ILC.CNR.IT.* Retrieved 08 07, 2021, from http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html: http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

Sitio Big Data. (2021, 08 04). *Construir un dataset en aprendizaje automático*. Retrieved from SITIOBIGDATA: https://sitiobigdata.com/2019/12/24/analisis-de-sentimientos-en-aprendizaje-automatico/

Torruella, J., & Llisterri, J. (1999). *“Diseño de corpus textuales y orales”, en J. M. Blecua, G. Clavería, C. Sánchez y J. Torruella (eds.): Filología e informática. Nuevas tecnologías en los estudios filológicos.* (D. d. Universidad Autónoma de Barcelona, Ed.) Retrieved 08 08, 2021, from http://liceu.uab.es/~joaquim/publicacions/Torruella_Llisterri_99.pdf

Veron, E. (1987). *La semiosis social. Fragmentos de una teoría de la discursividad.* Buenos Aires, Argentina: Gedissa.

Wallis, S. & Nelson G. (2001). *'Knowledge discovery in grammatically analysed corpora'. Data Mining and Knowledge Discovery.*

Yániz, C. y Villardón Gallego, L. . (2006). *Planificar desde competencias para promover el aprendizaje: el reto de la sociedad del conocimiento para el profesorado universitario.* Bilbao: Publicaciones de la Universidad de Deusto.

Yu, HF., Huang, FL. & Lin, CJ. . (2011). *Dual coordinate descent methods for logistic regression and maximum entropy models.* (Vol. 85). Machine Learning Journal.

ANEXOS

Anexo A: Encuesta

ENCUESTA

Objetivo: Obtener un diagnóstico sobre usabilidad del Corpus**.**

Marque con una “X” la respuesta que usted considere apropiada, en las preguntas enunciadas a continuación del 1 al 4.

1. Evalúe el grado de funcionalidad de la herramienta
1. Pobre _____
2. Regular _____
3. Bueno _____
4. Excelente _____

2. ¿Le gustó cómo la herramienta manejó la información?
1. Fue caótico _____
2. Fue confuso _____
3. Fue ordenado _____
4. Definitivamente me gustó _____

3. ¿Qué tan precisa es la herramienta?
1. Muy imprecisa _____
2. Imprecisa_____
3. Precisa _____
4. Muy precisa _____

4. En general, ¿qué tan fácil fue usar la herramienta?
1. Muy difícil _____
2. Difícil_____
3. Fácil _____
4. Muy fácil _____