BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

Facultad de Ciencias de la Computación
Maestría en Ciencias de la Computación

Aprendizaje Ontológico para el Dominio de Desórdenes del Habla

Tesis presentada para obtener el grado de Maestría en Ciencias de la Computación

Presenta:
LCC CONCEPCIÓN STEPHANIE VÁZQUEZ GONZÁLEZ

Directora de tesis:
Dr. MARÍA JOSEFA SOMODEVILLA GARCÍA

Asesor de tesis:
Dr. IVO H. PINEDA TORRES

Octubre 2018

Índice General

Resumen

El presente proyecto consiste en la realización de una Ontología que representa el dominio de los desórdenes del habla en niños con la finalidad de ser una herramienta de soporte a los terapeutas para el diagnóstico y posible tratamiento de los desórdenes antes mencionados.

Los desórdenes del habla serán clasificados utilizando una taxonomía obtenida de un corpus de desórdenes del habla previamente conformado utilizando técnicas de Procesamiento de Lenguaje Natural (PLN) y Recuperación de Información (RI). Basada en esta taxonomía, la ontología, la cual estructura y formaliza conceptos definidos por los principales autores del tema, es desarrollada. Las clases principales de la ontología representan la clasificación taxonómica de los desórdenes del habla, su origen etiológico, síntomas y signos de cada desorden, y estrategias de evaluación e intervención; también están representados los pacientes y terapeutas como instancias. La importancia de una detección y diagnóstico temprano de un desorden del habla -que puede tener un impacto social, económico y educativo, radica en que el pronóstico del tratamiento depende de la causa del trastorno y de un tratamiento oportuno.

Capítulo 1

Descripción del Proyecto

1.1 Introducción

En el área de la Inteligencia Artificial, el conocimiento en un sistema computacional es pensado como algo que está representado explícitamente y operado por procesos de inferencia. Sin embargo, esa es una visión muy estrecha. Cualquier software que realice una tarea útil no puede ser desarrollado sin una relación a un modelo del mundo relevante -entidades, propiedades y relaciones en ese mundo-. Los sistemas de recuperación de información, las bibliotecas digitales, la integración de fuentes de información heterogéneas e incluso los motores de búsqueda por internet necesitan ontologías de dominio para organizar la información y dirigir los procesos de búsqueda [1].

En los sistemas de información al modelar grandes dominios del conocimiento, las ontologías de dominio se volverán tan importantes en el software en general como en las diferentes áreas de la IA. En la IA mientras el conocimiento abarca al campo entero hay dos áreas de aplicación en particular que dependen de un cuerpo rico de conocimiento. Una de ellas es el Procesamiento del Lenguaje Natural, las ontologías son útiles para en PLN en dos maneras. Primero, el dominio de conocimiento juega un rol crucial en la desambiguación, una ontología de dominio bien diseñada provee la base para la representación del dominio de conocimiento. Además, una ontología de dominio ayuda a identificar las categorías semánticas que están relacionadas en el entendimiento del dominio de discurso. Para este uso, la ontología juega el rol de un diccionario de conceptos. En general en PLN se necesitan tanto una ontología superior de propósito general y una ontología de dominio específico que se enfoque en el dominio de discurso. CYC [2], Wordnet [3]y Sensus son ejemplos de ontologías de propósito general compartibles que son utilizadas para comprensión del lenguaje.

La resolución se problemas basada en conocimiento (KBPS) es la segunda área de la IA que emplea ampliamente el conocimiento. Estos sistemas resuelven una variedad de problemas -tales como diagnóstico, planeamiento y diseño- utilizando una base rica de conocimiento. Actualmente estos sistemas emplean conocimiento específico del dominio, el cual es a menudo suficiente para construcción de sistemas de conocimiento que apuntan a áreas específicas de aplicación. Pero se ha propuesto que se complemente este tipo de sistemas con conocimiento de sentido común en adición a el conocimiento de dominio específico. Con las ontologías podemos construir bases de conocimiento utilizando la estructura de conceptualización para codificar piezas específicas de conocimiento. Así estas bases de conocimiento

pueden ser compartidas de forma más confiable, y ayudar a clarificar la representación semántica.

1.2 Antecedentes del Proyecto

Un desorden del habla es la dificultad de producir o formar los sonidos específicos del habla para comunicarse. Estos desórdenes pueden ser desde simples sustituciones de sonidos hasta la inhabilidad de entender o utilizar el lenguaje o mecanismo motor-oral para el habla. Sus causas son tan diversas como la pérdida auditiva, trastornos neurológicos, lesiones cerebrales, discapacidades intelectuales, o impedimentos físicos como labio leporino [4].

Según estadísticas de Global Disability Rights el 7.5% de la población total en México (año 2015) padece de alguna discapacidad (aproximadamente 9.17 millones de personas) y de esas el 4.87% padece algún desorden del habla (0.45 millones de personas) [5]. En los niños y jóvenes, las discapacidades para hablar tienen un lugar importante, en algunos casos 2 y 4 veces más altas que en los adultos.

Un desorden del habla puede tener repercusiones económicas y de aprendizaje, en cuanto a las económicas como algunas otras discapacidades está ligada de forma bilateral a la pobreza, ya que una discapacidad puede aumentar el riesgo de pobreza y la pobreza puede aumentar el riesgo de una discapacidad. Una discapacidad puede empeorar el bienestar social y económico a través de diferentes canales, como son un impacto adverso en la educación, en el empleo, ingresos, y aumentar los gastos debido a la discapacidad[6].

Los niños con algún desorden del habla tienen una menor posibilidad de asistir a la escuela lo cual limita las oportunidades de formación de capital humano y afronta una oportunidad de empleo reducida y una productividad reducida como adulto. Así mismo las personas con alguna discapacidad son más propensas a estar desempleadas y generalmente perciben un menor ingreso cuando se encuentran empleadas[7].

Los desórdenes del habla también tienen el potencial de aislar a los individuos del medio social razón por la cual es necesaria una intervención apropiada del problema. Los desórdenes no tratados pueden convertirse en un trastorno causante de dificultades en el aprendizaje, ya que las destrezas del lenguaje son más fáciles de aprender antes de los 5 años[4].

Los pasos que se llevan a cabo para la identificación y tratamiento de un desorden del habla incluyen:

- Identificación del niño con el desorden o impedimento del habla,
- Diagnóstico y valoración del desorden específico,
- Remisión para atención profesional necesaria,
- Provisión de servicios de terapia de lenguaje para la mejora y prevención de impedimentos de comunicación o del habla, y
- Consejería o guía para los padres, niños y educadores con respecto a los desórdenes del habla.

Las Tecnologías de la Información y Comunicación (TIC) puede asistir a los terapeutas y pacientes en la mayoría de estos pasos de diagnóstico y tratamiento en niños incluso en los casos de discapacidades severas para proporcionar el cuidado adecuado.

Al manipular una cantidad grande de información como el caso de los desórdenes del habla y enfrentarse al problema de su organización, una manera de resolver tal problema es con el uso de una ontología la cual provee a través de la Web Semántica la información y servicios para satisfacer las necesidades de contenido de los usuarios e incluso las generadas por máquinas [8]. Las ontologías aportan una estructura bien definida y no ambigua para una representación clara y precisa de información concerniente a un dominio en particular, como en este caso los desórdenes del habla, y así, ser una herramienta para el diagnóstico de tales trastornos. Las ontologías están conformadas de dos componentes principales: clases y relaciones (Ver Fig. 1)

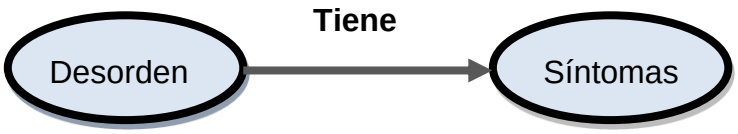

Fig. 1 Representación simple de los dos componentes principales en una ontología: clases y relaciones.

La ontología es propuesta para organizar y buscar información relativa a los desórdenes del habla tales como diferentes desórdenes, características de cada desorden, planes de terapia, taxonomía de los desórdenes del habla, y alguna otra información de ayuda para los terapeutas y pacientes, así como también las relaciones entre todos ellos.

La importancia de las ontologías radica en que el análisis ontológico clarifica la estructura de conocimiento, dado un dominio, su ontología es el centro de cualquier sistema de representación de conocimiento para ese dominio. De este modo, el primer paso al construir un sistema efectivo de representación del conocimiento y un vocabulario es realizar un análisis ontológico efectivo del campo o dominio.

Otro punto importante es que las ontologías permiten compartir el conocimiento. La ontología captura la estructura conceptual intrínseca del dominio. A fin de construir un lenguaje de representación del conocimiento basado en el análisis, se necesitan asociar términos con los conceptos y relaciones en la ontología y concebir una sintaxis para codificación del conocimiento en términos de los conceptos y relaciones. Este lenguaje de representación del conocimiento se puede compartir con otros que tengan necesidades similares de representación del conocimiento en ese dominio, y elimina así la necesidad de replicación del proceso de análisis del conocimiento [1].

Uno de los primeros pasos en el desarrollo de esta ontología es la conformación de un Corpus, en este caso de documentos relacionados al dominio de los desórdenes del habla. Un Corpus es una gran colección de textos. Es un cuerpo de material escrito o hablado sobre el

cual está basado un análisis lingüístico. El análisis del Corpus provee información léxica, morfosintáctica, semántica y pragmática [9].

Previamente ya han sido utilizadas ontologías en el dominio de *e-health* para modelar almacenes de datos clínicos debido a que esta estrategia es suficientemente flexible y eficiente para la complejidad y el constante cambio en el dominio de temas médicos y minimiza a su vez el conocimiento técnico requerido por los usuarios finales [10].

1.3 Hipótesis, Objetivos y Preguntas de investigación

Con el presente trabajo se planea construir una herramienta auxiliar para diagnóstico de desórdenes del habla en niños, ya sea en diagnóstico presencial o remoto. Partiendo de la hipótesis de que el uso de TICs y técnicas de PLN pueden asistir en el trabajo a los terapeutas en el diagnóstico de un desorden del habla. La ontología resultante será útil para la organización y representación de toda la información referente al dominio y desplegarla según las necesidades del terapeuta. Esto ayudará a hacer más diagnósticos oportunos dentro de las escuelas primarias donde se refieran al terapeuta a niños con un posible desorden que podría afectar su aprendizaje y futuro desarrollo.

El objetivo general es desarrollar una ontología del dominio de los desórdenes del habla como herramienta de soporte a terapeutas para toma de decisiones sobre diagnóstico y terapia. Esto abarca desde la creación de un corpus referente a desórdenes del habla hasta la fase de poblado de la ontología y comparación de resultados. Así podemos entonces desglosar los siguientes:

Objetivos específicos

1. Recopilar información sobre desórdenes del habla a través de la revisión de la literatura, así como de aplicaciones y software existente referente al tema.
2. Crear un corpus sobre DdH utilizando documentos recuperados de la Web referentes a desórdenes del habla en niños.
3. Diseñar una taxonomía utilizando el corpus de documentos sobre DdH empleando técnicas de PLN y RI.
4. Desarrollar una ontología basada en la taxonomía, partiendo del planteamiento de las preguntas de competencia hechas por terapeutas e instructores para diagnóstico y tratamiento. Identificar las propiedades de las clases y las relaciones entre las clases para formalizar la ontología.
5. Formalizar las preguntas de competencia utilizando lógica de primer orden.
6. Implementar las relaciones estructurales y de comportamiento en Protégé.
7. Evaluar la consistencia y desempeño de la ontología a través de la ejecución del razonador, evaluación de las preguntas de competencia y consulta con expertos del área.

8. Poblado de la ontología con datos obtenidos de pacientes y terapeutas.
9. Comparar resultados con otras ontologías y aplicaciones similares, así como prueba con usuarios para retroalimentación.

Analizando los objetivos anteriores a continuación se proponen las preguntas de investigación, iniciando con una pregunta general: ¿Es posible desarrollar una herramienta basada en aprendizaje ontológico que sirva como auxiliar en el tratamiento de los desórdenes de habla en niños de edad escolar?

Siendo las preguntas siguientes relacionadas con los objetivos específicos:

- ¿Es posible generar un corpus sobre desórdenes del habla recuperando documentos de la Web?
- Una taxonomía generada con base en ese corpus ¿es correcta?
- ¿Puede una ontología basada en dicha taxonomía responder a las necesidades de toma de decisión de un terapeuta del lenguaje?
- ¿El tiempo de respuesta de la ontología es aceptable durante el proceso de diagnóstico de un niño?
- ¿Puede la ontología ser una herramienta de ayuda también para el paciente y sus familiares?

1.4 Metodología

Para alcanzar los objetivos descritos anteriormente se propone una metodología consistente en 9 pasos para alcanzar cada uno de los objetivos, desde la recopilación de información referente al dominio, pasando por el diseño y desarrollo de las taxonomías y la ontología, hasta llegar a la evaluación de la herramienta obtenida.

1.5 Justificación

La detección de desórdenes de habla entre niños que asisten a educación básica se da en primera instancia por parte del profesor de grupo y con retroalimentación de los padres del infante sobre la situación que presenta el niño con deficiencias de pronunciación. Es el profesor el que generalmente solicita la evaluación y opinión de personal capacitado en el tema para la confirmación del diagnóstico si es el caso. Para los terapeutas el tener a mano las bases de conocimiento de los desórdenes del habla es de suma importancia para detectar uno o más desórdenes en un niño o en su caso para dar un diagnóstico negativo a la detección de un desorden real y sólo se trate de falta de madurez en el niño. Siendo reducido el número de terapeutas dentro de las escuelas públicas la detección y posterior tratamiento de los niños suele ser más tardado que el tiempo que debiera ser el óptimo para la atención de estos trastornos. Un mismo terapeuta suele atender diferentes escuelas durante sus horas de trabajo y

hacer sólo un diagnóstico inicial presencial, la posterior terapia se hace en casa trabajando en conjunto con los padres y generalmente no se lleva un registro de los pacientes ni de los avances registrados. El proveer a los terapeutas con una herramienta que agilice su trabajo, tenga datos sobre los pacientes contenidos en el poblado y proponga soluciones para la terapia es de gran ayuda para acelerar la detección de estos desórdenes y poner al alcance de padres, maestros, terapeutas y pacientes la información pertinente para cada etapa del diagnóstico y tratamiento del niño.

Así mismo, la creación de una ontología sobre los desórdenes del habla en español servirá de base para futuros usos del conocimiento representado en dicha ontología al ser compartido, expandido o reutilizado.

1.6 Resultados Esperados

Se espera obtener una ontología que responda a las preguntas de competencia de manera correcta y que cumpla con las necesidades de información del usuario final con tiempos de respuesta aceptable en comparación con la literatura y estado del arte consultados en este documento.

Impacto Socioeconómico

La importancia de una detección y diagnóstico temprano de un desorden del habla radica en el impacto social, económico y educativo que tienen dichos desórdenes en la vida del niño.

Las personas con algún tipo de discapacidad tienen peores condiciones de salud, menor nivel educativo, menor participación económica e índices más altos de pobreza en comparación con personas sin discapacidades. Esto en parte es porque la gente con algún tipo de discapacidad experimenta barreras al acceder a servicios que la mayoría da por sentado, incluyendo salud, educación, empleo, y transporte, así como también el acceso a información, dicha dificultades se ven ampliadas en comunidades pobres. Esto causa que no reciban los servicios requeridos para tratamiento de su discapacidad, y experimentan exclusión en las actividades de la vida diaria.

1.7 Límites del presente trabajo

Los límites de la tesis radican en la creación manual de la ontología basada en documentos recuperados de la web que resulten relevantes al dominio de los desórdenes del habla. Al recuperar una gran cantidad de documentos relevantes al tema se encontró que la mayoría estaban en idioma inglés por lo cual algunas de las técnicas de recuperación de información tuvieron que realizarse utilizando este idioma para la búsqueda de términos relevantes y relaciones entre ellos, además de que existen pequeñas variaciones en la teoría de los trastornos de habla en cada idioma en particular por lo tanto se debe adaptar o descartar algunos

de los documentos recuperados para cumplir con el objetivo de la tesis. Se genera una herramienta que pueda ser aplicada en el ámbito de la terapia de desórdenes como auxiliar para los terapeutas al momento del diagnóstico y para brindar información sobre los desórdenes más no una herramienta de terapia. El trabajo también está enfocado sólo a la parte de desórdenes del habla en idioma español y no del lenguaje en general.

1.8 Aportaciones

La aportación de la presente investigación radica principalmente en la creación de una taxonomía más completa para la formalización de conceptos de la ontología de desórdenes del habla. En el ámbito computacional se plantea una metodología para el desarrollo de la ontología basada en un corpus de documentos relevantes y la generación de una ontología que pueda ser utilizada a futuro para su ampliación o inclusión dentro de una ontología que abarque más desórdenes relacionados al lenguaje.

1.9 Contenido de la tesis

El presente trabajo está estructurado de la manera siguiente: el capítulo 2 presenta las bases teóricas y el estado del arte tanto de los desórdenes del habla como de las tecnologías aplicadas a estos desórdenes, en particular las ontologías. El capítulo 3 explica la construcción de los recursos léxicos necesarios para el desarrollo de la ontología. El capítulo 4 habla sobre el desarrollo de la ontología en la herramienta Protégé. El capítulo 5 presenta los resultados y conclusiones obtenidos y el capítulo 6 menciona el trabajo a futuro que podría hacerse con la presente tesis como base.

Capítulo 2

Marco Teórico

2.1 Estado del Campo o del Arte

Comencemos con una definición de los conceptos más importantes para el proyecto, referentes al dominio del conocimiento de interés y a las tecnologías aplicadas a él.

1. Desórdenes del habla

El ***habla*** es la producción expresiva de sonidos e incluye la articulación individual, fluidez, voz y resonancia [11].

Un ***desorden del habla*** según Bashir es un término que representa un grupo heterogéneo de discapacidades ya sea del desarrollo o adquiridas caracterizadas principalmente por déficits de comprensión, producción y/o uso del lenguaje. Los desórdenes del lenguaje son crónicos y pueden durar a través de toda la vida del individuo [12].

Los desórdenes de comunicación incluyen a los del lenguaje, del sonido del habla (fonológico), de la comunicación social (pragmático) y de la fluidez de inicio en la infancia (tartamudeo). Los tres primeros se caracterizan por déficits en el desarrollo y uso del lenguaje, habla, y comunicación social, respectivamente. El desorden de la fluidez de inicio en la infancia está caracterizado por perturbaciones en la fluidez natural y producción motora del habla, incluye el sonido repetitivo de sílabas, prolongación del sonido de consonantes o vocales, palabras truncas, bloqueo, o palabras producidas con un exceso de tensión física. Como otros desórdenes del desarrollo neurológico, los desórdenes de comunicación comienzan a temprana edad y pueden producir discapacidades funcionales de por vida.

Los desórdenes de habla están incluidos dentro de los desórdenes de la comunicación y estos a su vez dentro de los desórdenes del desarrollo neurológico en la clasificación del DSM-5 (*Diagnostic and Statistical Manual of Mental Disorders, Quinta Edición*) [11]. El DSM-5 es un manual realizado por la Asociación Americana de Psiquiatría (*APA*) que contiene descripciones, síntomas y criterios varios para el diagnóstico de todos los trastornos considerados mentales. Estos criterios de diagnóstico nos proporcionan un lenguaje común para una clasificación precisa y consistente de los desórdenes del habla. El DSM en sus diferentes versiones ha sido el sistema de clasificación más aceptado tanto para diagnósticos como para investigación y docencia desde su primera publicación en 1952.

Según la clasificación de Gallardo y Gallego los Trastornos del Lenguaje y Comunicación incluyen los Trastornos del Lenguaje Verbal y a su vez se dividen en Trastornos del Habla y Trastornos del Lenguaje [13]. Siendo los Trastornos del Habla los que forman el dominio de interés para la ontología. Esta clasificación está representada en la Figura 2 a continuación.

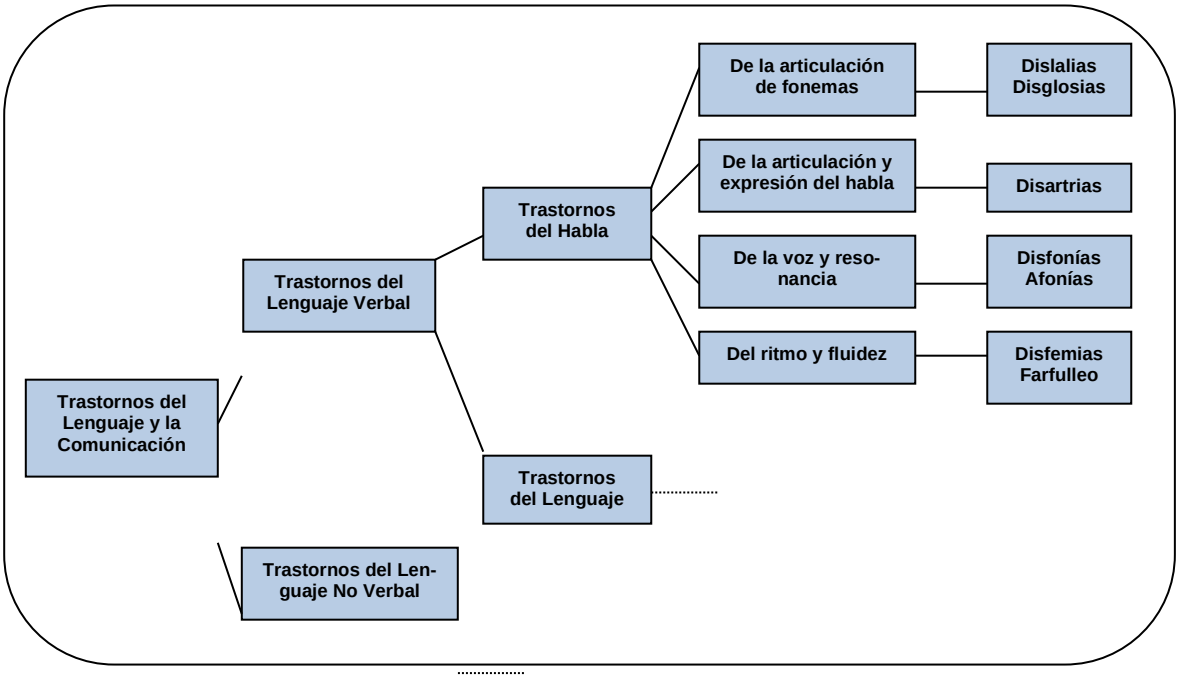

Fig. 2 Clasificación de Trastornos del Lenguaje por Gallardo y Gallego.

Esta clasificación nos presenta los principales trastornos del habla en niños y que son los que serán incluidos dentro de la ontología: **dislalia, disglosia, disartria, disfonía, afonía, trastorno inconsciente, disfemia** y **farfulleo**.

Una definición rápida de cada una es incluida a continuación:

Dislalia: es un trastorno de la articulación de los fonemas. Se trata de una incapacidad para pronunciar correctamente ciertos fonemas o grupos de fonemas. Es debida a un mal uso de los órganos articulatorios. Es el trastorno del lenguaje más común en los niños, el más conocido y más fácil de identificar. Suele presentarse entre los tres y los cinco años, con alteraciones en la articulación de los fonemas.

Gallardo y Gallego clasifican las dislalias de la siguiente manera:

- *Evolutiva o fisiológica*: El sujeto no ha adquirido una madurez en su aparato fono-articulatorio, no articula o distorsiona algunos fonemas, pero desaparecen con el tiempo. Se precisa un posible programa de prevención de alteraciones del lenguaje enfocado hacia la respiración, el soplo, los movimientos de labios y lengua.

- *Audiógena*: Originada por una deficiencia auditiva, el niño no oye bien y, por tanto, no articula correctamente, confundiendo fonemas. Se evidencian errores en la voz. Se precisa de un programa de intervención para el desarrollo de la percepción auditiva.
- *Orgánica*: Se presenta un problema orgánico en el SNC o en los órganos fono-articuladores.
    a. Disartrias (alteraciones en centros neurales cerebrales)
    b. Diglosias (anomalías en órganos del habla: labios, lengua)
- *Funcional*: Ocasionada por un mal funcionamiento de los órganos articulatorios, aunque no se evidencia daño ni lesión orgánica.

Disglosia: como ya se mencionó anteriormente es lo mismo que una dislalia orgánica, es un trastorno de la articulación de fonemas por alteración o daño de los órganos periféricos del habla, un trastorno provocado por lecciones o malformaciones de órganos articulatorios periféricos y no de origen neural central.

Disartria: es el nombre que se da a una serie de trastornos motores del habla que surgen como resultado de un daño en el sistema nervioso y que se manifiestan por dificultades neuromusculares.

Disfonía: es la pérdida parcial de la fonación o alteración de la voz en cualquiera de sus cualidades, a causa de de un trastorno orgánico o por su inadecuada utilización.

Afonía: es la pérdida total de la voz a causa de un estado inflamatorio agudo, un traumatismo, parálisis laríngeas, cuadro psíquico histérico, etc.

Trastorno inconsistente: lo característico de este trastorno es la variación. En todos los niños, incluidos los que muestran un desarrollo típico, se observan variaciones en la pronunciación de muchas palabras o de rasgos fonológicos, como los grupos consonánticos [14].

Disfemia: también conocido como tartamudeo se trata de un trastorno del ritmo del habla, con bloqueos, tics, repeticiones o prolongaciones de sonidos que dificultan la fluidez.

Farfulleo: es un trastorno en la fluidez y ritmo verbal, que se caracteriza por taquilalia y falta de inteligibilidad. La persona habla demasiado rápido, lo que produce distorsiones en el ritmo y la articulación. El lenguaje es errático, confuso, entrecortado y suele incluir patrones gramaticales erróneos.

Al diagnosticar cualquier enfermedad, padecimiento, síndrome, trastorno, desorden etc., se debe hacer una exclusión de otras posibles causas con un cuadro clínico semejante y que no son el diagnóstico final, este proceso es llamado Diagnóstico diferencial. [15]

Se debe tomar en cuenta lo siguiente antes de realizar el diagnóstico:

- Variaciones normarles del lenguaje: variaciones regionales, sociales y culturales del lenguaje.
- Deficiencias de audición o sensoriales: sordera que conlleve anormalidades del habla.
- Deficiencias estructurales: paladar hendido, etc. pueden causar una alteración del habla.

- Disartria: alteraciones debidas a un trastorno motor como parálisis cerebral.
- Mutismo selectivo: debido a un trastorno de ansiedad caracterizado por ausencia del habla en uno o más contextos o entornos. El mutismo selectivo se puede manifestar en niños con algún trastorno del habla por vergüenza a causa de sus deficiencias, aunque muchos niños con mutismo selectivo muestran un habla normal en entornos "seguros".
- Déficit sensitivo: una deficiencia auditiva u otro déficit sensitivo o motor del habla puede asociarse con una disfluencia del habla.
- Disfluencia normal del habla: el trastorno debe distinguirse de las disfluencias normales presentes en niños pequeños. Si estas aumentan en frecuencia o complejidad al crecer el niño, se podría entonces ya pensar en un diagnóstico de desorden de la fluidez.
- Efectos secundarios de medicación: el tartamudeo puede ser a causa de un efecto secundario de algún medicamento, puede detectarse por su relación temporal con la exposición al medicamento.
- Trastorno de Tourette: los tics vocales y vocalizaciones repetitivas del trastorno de Tourette deben poderse distinguir de los sonidos repetitivos del trastorno de la fluidez debido a su naturaleza y ritmo.

En cuanto a la terapia de un desorden del habla, la terapia del habla y del lenguaje es el tratamiento para la mayoría de los niños con discapacidades del habla y aprendizaje del lenguaje. Las discapacidades en el habla se refieren a problemas con la producción de sonidos, mientras que los problemas con el aprendizaje del lenguaje son las dificultades al combinar las palabras para expresar ideas.

La Asociación Americana del Habla, Lenguaje y Audición (*American Speech-Language-Hearing Association*, ASHA) clasifica los trastornos del habla según describimos a continuación:

- Los trastornos de articulación - dificultad producir sonidos en las silabas y al emitir palabras de forma incorrecta de modo que otras personas no pueden entender lo que la persona está diciendo.
- Trastornos con la fluidez del habla con problemas que incluyen tartamudez - una condición donde el habla se interrumpe debido a pausas anormales, repeticiones o sonidos prolongados y silabas.
- Resonancia o trastornos de la voz - incluye problemas con el tono, el volumen o la calidad de la voz. Distrae a los oyentes de lo que se está diciendo. Estos tipos de trastornos también pueden causar dolor al niño o hacerle sentir incómodo cuando está hablando.
- Disfagia oral/trastornos de la alimentación - incluye dificultades al comer o al tragar.

Los trastornos del lenguaje pueden ser receptivos o expresivos:

- Los trastornos receptivos se refieren a las dificultades al entender o procesar el lenguaje.
- Los trastornos expresivos incluyen dificultades para combinar palabras, vocabulario limitado o inhabilidad de usar el lenguaje en forma socialmente apropiada.

Los fonoaudiólogos o logopedas, generalmente conocidos como terapeutas del habla, son profesionales educados en el estudio de la comunicación humana, su desarrollo y sus trastornos. Al evaluar las habilidades del habla, lenguaje, comunicación cognitiva y la forma de tragar de los niños y adultos, los patólogos del habla y del lenguaje pueden identificar problemas en la comunicación y la mejor manera de tratarlos.
Los terapeutas del habla atienden los trastornos en la articulación del lenguaje, problemas con su fluidez, trastornos orales, motores y de la voz, así como trastornos en el lenguaje receptivo y expresivo.

Generalmente el terapeuta apropiado trabajará con el niño individualmente, en un pequeño grupo o directamente en un aula de clase para sobrellevar las dificultades que incluye cada trastorno en particular.
Los terapeutas utilizan una variedad de estrategias incluyendo: **Actividades de intervención del lenguaje.** En estos ejercicios el Patólogo del Habla y del Lenguaje interactuará con un niño jugando y hablando. **Terapia de la articulación.** Los ejercicios de articulación o producción de los sonidos incluyen la pronunciación correcta de sonidos y silabas por parte del terapeuta generalmente durante actividades de juego. **Terapia oral y motora de la alimentación.** El terapeuta utilizará una variedad de ejercicios, incluyendo el masaje facial, y movimientos para ejercitar la lengua, labios y mandíbula que fortalecen los músculos de la boca [16].

2. Ontologías e ICT aplicadas a los desórdenes del habla

Ahora que se han revisado los desórdenes del habla que nos incumben dentro del dominio de la ontología a realizar pasaremos a revisar algunos conceptos relacionados con el diseño de la ontología misma.

Ya se mencionó anteriormente a la **Web Semántica**, veamos una definición al respecto; Tim Berners Lee define la define de la siguiente manera: La Web Semántica no es una Web separada sino una extensión de la actual, en la cual a la información le es dado un significado bien definido, y permite un mejor trabajo cooperativo entre computadoras y personas [17].

Un pilar importante para la web semántica son las ontologías, las cuales proveen una estructura bien definida para presentar información acerca de un dominio en particular de una forma clara y precisa. Según Thomas R. Gruber: Una **ontología** es una especificación explícita de una conceptualización, se describe la ontología de un programa al definir un conjunto de término representacionales. Las definiciones asocian los nombres de entidades

en el universo del discurso (clases, relaciones, funciones, u otros objetos) con texto humanamente legible y describen lo que denotan los nombres, y axiomas formales que limitan la interpretación y uso de esos términos [18].

En su trabajo, Adolfo Lozano Tello menciona que los principales elementos de las ontologías son los siguientes [19]:

- *Conceptos*: son las ideas básicas que se intentan formalizar, es decir el conocimiento del tema.
- *Relaciones:* representan la interacción y enlace entre los conceptos del dominio.
- *Funciones:* son un tipo de relación donde se obtiene un elemento mediante el cálculo que involucra varios elementos de la ontología.
- *Instancias:* se utilizan para representar objetos determinados de un concepto.
- *Axiomas:* son reglas o normas que se declaran sobre relaciones que deben cumplir los elementos de la ontología.

En 2009 un modelo propuesto para la evaluación de ontologías por E. Ramos, H Nuñez y R. Casañas; se menciona que los aspectos más importantes de una ontología son: el uso de un lenguaje correcto y legible para cada miembro de la audiencia de cada elemento de la ontología; exactitud de la estructura taxonómica, el uso de un lenguaje de codificación adecuado; su aplicabilidad o adecuación a los requerimientos [20]. Las tecnologías de la Web Semántica han demostrado ser realmente útiles en la vida diaria, aplicadas a situaciones más generales y nos dan un poder más directo que permite mejores resultados de búsqueda [21].

Varias metodologías para la construcción de ontologías han sido definidas, como lo mencionan Corcho, M. Fernández-López y A. Gomez-Pérez [22]. Varias de ellas pueden ser utilizadas para el desarrollo de la ontología de nuestro interés tomando elementos como los del enfoque sugerido por Grüninger y Fox [23], o algunos sugeridos por la metodología METHONTOLOGY [24]. La metodología de Grüninger y Fox's está basada en el diseño de sistemas con una base de conocimiento usando lógica de primer orden. Primero, el escenario en el cual la ontología es aplicable se define, seguido por la generación en lenguaje natural de las llamadas preguntas de competencia, cuyo objetivo es determinar el alcance de la ontología. Estas preguntas y sus respectivas respuestas son utilizadas para extraer los conceptos principales, así como sus relaciones, propiedades y axiomas dentro de la ontología. La formalidad de este método basado en lógica clásica nos permite transformas escenarios informales en modelos computacionales. Los elementos de METHONTOLOGY que podrías ser útiles para el diseño son los siguientes: construcción de una taxonomía para la base del conocimiento a ser representada; tablas de descripción de clases, de atributos y relaciones entre clases y de reglas y atributos de las instancias.

Posteriormente, al tener ya hecha la ontología se necesita también alguna metodología para evaluar desde el lenguaje utilizado para su codificación hasta la adecuación al escenario en el cual será utilizada [20].

Para evaluar el *uso correcto del lenguaje* se puede validar que el lenguaje cumpla con estándares para desarrollos ontológicos, como: OWL (Ontology Web Language), RDF (Resource Description Framework), DAML (DARPA Agent Markup Language), etc. utilizando el marco de prueba que provee el editor de ontologías Protégé-OWL.

En cuanto a la *exactitud de la estructura taxonómica* para examinar su rigurosidad que representa los conceptos, términos y clases del dominio, así como la naturaleza de las diferentes relaciones jerárquicas y semánticas, es necesario, y en algunas oportunidades imprescindible, el conocimiento que sólo los expertos en el área pueden proporcionar. La evaluación taxonómica considera el chequeo de inconsistencias, completitud y redundancia de los términos de la taxonomía. Los errores más comunes son: clasificaciones semánticas incorrectas (clasificación de conceptos como subclase de una clase a la que no pertenecen), clases e instancias con diferentes nombres, pero definiciones similares, omisión de conocimiento disjunto entre clases, ausencia de conceptos, redundancia de relaciones (clases con más de una relación de subclase), clases definidas como generalización o especializaciones de sí misma, entre otros.

La *validez del vocabulario* se realiza al checar que los términos codificados en la ontología existan y sean significativos en otras fuentes de conocimiento independientes, como por ejemplo, el conocimiento contenido en el corpus del dominio, entendiéndose por corpus, al conjunto más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación. En esta fase lo que se realiza en primera instancia es un análisis del corpus del dominio, es decir; identificar, extraer y organizar los términos significativos del dominio a partir de esos documentos. Para evaluar el vocabulario se deben considerar medias de calidad de los resultados con técnicas de RI como son la precisión y el recall (exhaustividad).

La precisión se refiere al porcentaje de los términos de la ontología que aparecen en el corpus con relación a la cantidad total de términos de ontología y se expresa de la siguiente manera:

$$\textbf{\textit{Precisión}} = CO\text{-}C \,/\, COnto$$

CO-C = Cantidad de términos que se solapan entre la ontología y el corpus.
COnto = Cantidad total de términos de la ontología.

Mientras que el recall se refiere al porcentaje de términos del corpus que aparecen en la ontología con relación al total de términos en el corpus y utiliza la siguiente expresión:

$$\textbf{\textit{Recall}} = CO\text{-}C \,/\, CCorp$$

CCorp = Cantidad total de términos del corpus

En función de los valores obtenidos para la Precisión y el Recall, se establece una valoración cualitativa acerca de lo adecuado del vocabulario.

Finalmente, la *adecuación a requerimientos* consiste en elaborar un reporte que especifique para qué construir una ontología, cuál es su propósito, una descripción del dominio de aplicación, sus posibles aplicaciones, nivel de formalidad, fuentes de conocimiento disponibles, sus usuarios potenciales y escenarios de uso. También se debe incluir las preguntas de competencia que serán utilizadas como un indicador del alcance y contenido del dominio representado.

Al hacer una revisión del estado del arte relativo a software y aplicaciones desarrollados en el dominio de los desórdenes del habla encontramos lo siguiente:

Dentro del área de la terapia del habla y lenguaje se han desarrollado trabajos que utilizan tecnologías de información y comunicación (TIC) enfocándose en algunos padecimientos específicos [25], para la clasificación automática de la calidad de pronunciación al atender desórdenes como la dislalia y disartria [26] o un sistema experto para la evaluación inicial de niños con posibles desórdenes del habla [27]. También podemos encontrar un llamado ecosistema de TIC inteligentes que incluyen administración de expedientes médicos electrónicos, vocabularios estandarizados, una base de datos de conocimiento, ontologías de conceptos del dominio del habla y lenguaje así como sistemas expertos enfocados en dar soporte a especialistas patólogos del habla y lenguaje, médicos, estudiantes, pacientes y a sus familiares[28]. Para la formación de profesionales en el área de los desórdenes del habla también existen herramientas basadas en ontologías y e-learning que dan soporte al entrenamiento y desarrollo de habilidades prácticas de futuros terapeutas del lenguaje al momento de diseñar planes de terapia [29]. Respecto a las terapias de lenguaje también se ha desarrollado una aplicación móvil que integra actividades de terapia para niños y utiliza lenguaje coloquial y juegos propios del estado de Chiapas [30]. Incluso existe una ontología que abarca varios aspectos de las terapias del habla y del lenguaje con conceptos como evaluación del paciente, pruebas realizadas, catálogo de médicos y terapeutas, listado de desórdenes, áreas referentes al habla y lenguaje, planes y ejercicios de terapia y seguimiento entre otros que utiliza ontologías y construcciones OpenEHR [31], [32]. Esta última tomándose como futura referencia para medición de los resultados obtenidos en la implementación de este proyecto.

En cuanto a la creación de taxonomías referentes a un dominio se han propuesto varios métodos que utilizan técnicas tan variadas como el análisis formal de conceptos y cláusulas de Horn con una validación hecha por inferencia lógica [33], clustering jerárquico de documentos basado en conjuntos de conceptos frecuentes y validado por la implementación de un prototipo [34] o un algoritmo generalizado de reglas de asociación que detecta relaciones entre conceptos y determina el nivel apropiado de abstracción al cual definir relaciones [35].

Referente a la construcción de un corpus las principales técnicas no han variado mucho con el tiempo, y los textos dentro del corpus necesariamente deben estar en forma electrónica. Entonces, la manera más práctica de construir un corpus es recopilar información que ya se encuentra digitalizada o hacer uso principalmente de transcripciones en forma electrónica de audios, o documentos [36].

Los trabajos antes mencionados tienen algunas limitantes debido a que algunos de ellos no están pensados en usuarios menos especializados como son profesores de educación básica, o las taxonomías y ontologías están enfocadas en un desorden muy específico o sólo tienen como objetivo la parte de la terapia de repetición dejando a un lado todo el proceso de diagnóstico.

Capítulo 3

Construcción de recursos léxicos

Uno de los primeros pasos para la implementación de la ontología es un método para la construcción y validación del corpus, en este caso, de documentos relativos al dominio de los desórdenes del habla. El método propuesto tiene la flexibilidad de retroalimentarse a sí mismo; una vez que un diccionario inicial es definido este puede ser actualizado con un diccionario extendido obtenido después de completar los pasos de este método.

Un corpus es una colección grande de textos. Es un cuerpo de material escrito o hablado sobre el cual se basa un análisis lingüístico. El corpus puede estar compuesto de lenguaje escrito, lenguaje hablado o ambos. Un corpus también puede ser abierto o cerrado. Un corpus abierto es aquel que no asegura contener toda la información de un área en específico mientras que un corpus cerrado asegura contener toda o casi toda la información de un campo en particular. Un *corpora* que puede ser procesable por computadora permite a los lingüistas adoptar el principio de responsabilidad total, recuperando todas las ocurrencias de una palabra o estructura particular para inspección o ejemplos seleccionados aleatoriamente. El análisis de corpus proporciona información léxica, morfosintáctica, semántica y pragmática [9].

3.1 Modelo de recuperación de información para la definición de los recursos léxicos

En orden de construir un corpus, es necesario recolectar una gran cantidad de documentos relevantes a los desórdenes del habla a través de un Web Crawler. Este crawler utiliza un diccionario primario con algunos de los términos relevantes al dominio, una vez que una cantidad representativa de estos documentos es obtenida, es necesario pre-procesarlos en varios pasos para limpiar y estandarizar la información a través de algoritmos de normalización y lematización. Una vez que los datos están limpios y normalizados para recuperar información algunos algoritmos como *Word frequency* y *n-gramas* son aplicados para extender el diccionario original relevante al dominio de desórdenes del habla.

Los distintos pasos para la construcción y procesamiento del corpus pueden ser vistos en el diagrama de la Figura 3.

3.2 Creación del corpus

La construcción de un corpus se divide en dos etapas: diseño e implementación. Una buena práctica en la etapa de diseño es definir qué tendrá el corpus idealmente, en términos de cantidad y tipo de lenguaje, y entonces los parámetros pueden ser ajustados mientras se lleva

a cabo la construcción, manteniendo un registro de qué está en el corpus, pudiendo así ser corregidos y añadidos posteriormente los documentos, y si otros utilizan el corpus sabrán qué contiene [36].

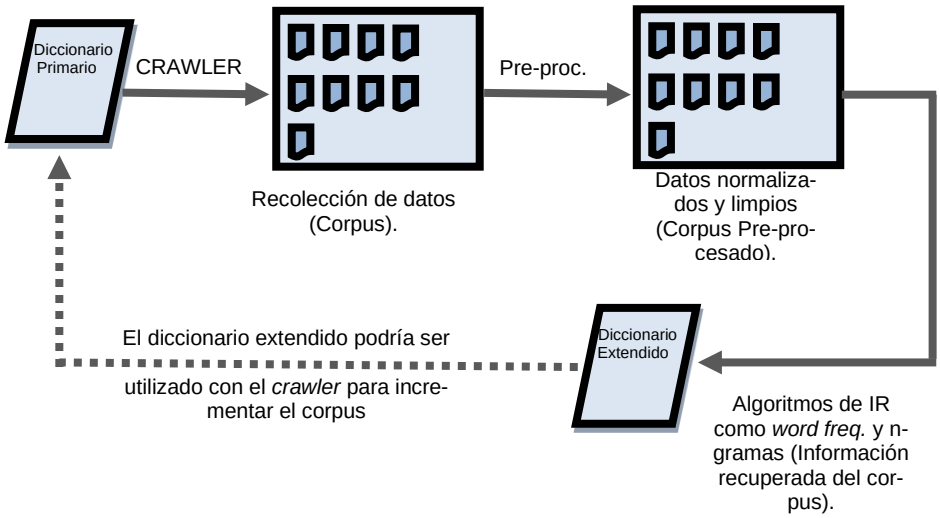

Fig. 3 Diagrama de pasos para la construcción y procesamiento del corpus.

Para construir un corpus existe un número de factores que deben ser tomados en consideración. Estos incluyen el tamaño, balance y representatividad. El tamaño del corpus depende mucho del tipo del tipo de preguntas que van a ser respondidas por él. Los documentos muestra necesitan estar balanceados con respecto a las respuestas de las preguntas de competencia. Obtener este balance de manera correcta según Wynne no es una ciencia exacta y no existe una forma confiable de determinar cuándo un corpus está realmente balanceado. Un acercamiento a obtener dicho balance es utilizar un corpus existente como modelo [36].

Se puede decir que un corpus es representativo si los hallazgos del corpus son generalizables a un lenguaje o a un aspecto particular del lenguaje como un todo. La noción de "saturación" podría ser útil en este caso. La saturación (a nivel léxico) se puede probar al tomar un corpus y dividirlos en secciones iguales en términos de número de palabras. Si otra sección del mismo tamaño es añadida, el número de elementos nuevos en la nueva sección debería ser aproximadamente el mismo que en las otras secciones [37].

La herramienta principal para recolectar la información para construir el corpus es un Web Crawler. Un crawler puede ser definido como un *bot* de internet que navega la World Wide Web, típicamente con el propósito de indexarla. Este crawler es alimentado con algunas páginas *semilla* para iniciar su tarea. En su núcleo es un elemento de recursión. El crawler debe recuperar contenidos de páginas desde un URL, examinar la página en búsqueda de otro URL, y recuperar tal página, ad infinitum [38]. Para encontrar documentos relevantes al dominio, y no solo una lista de links y datos aleatorios contenidos en las páginas semilla, es necesario establecer un diccionario primario al inicio del crawling.

3.2.1 Creación del diccionario

Este diccionario está conformado con algunas de las palabras más significativas dentro del dominio. Una manera simple de identificar estas palabras es tomar la taxonomía del dominio como base para crear tal lista de palabras y con la ayuda de expertos en el área. La figura 4 muestra la taxonomía de desórdenes del habla propuesta por el manual DSM-5 de desórdenes mentales [11].

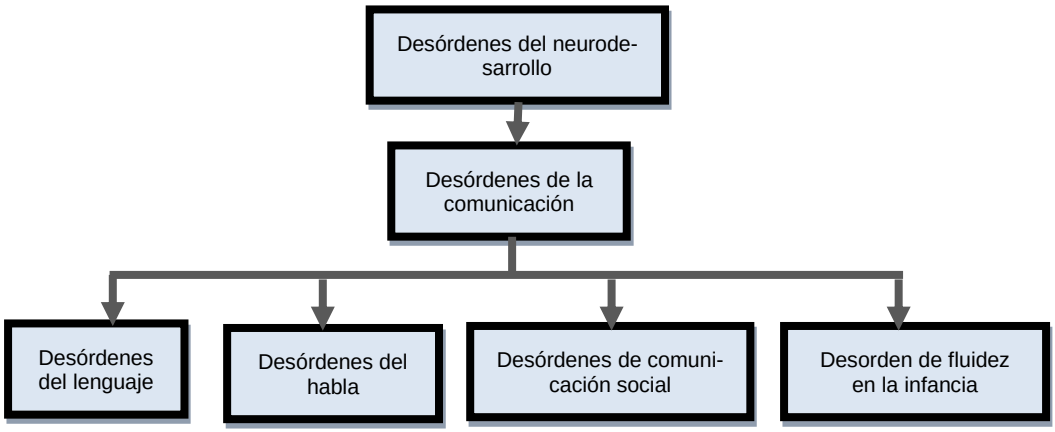

Fig. 4 Taxonomía jerárquica de Desórdenes del Habla de acuerdo al manual DSM-5.

Utilizando los términos del área de interés a la ontología de la anterior taxonomía podemos comenzar con la construcción del diccionario primario para acotar los resultados del crawler. Como la anterior (Figura 4) es una taxonomía pequeña y sólo nos interesa una rama terminal de ella, el tamaño del diccionario lo incrementaremos utilizando algunos otros términos relacionados. Existen algunas otras clasificaciones para los desórdenes del habla que incluyen nombres específicos para cada tipo de desorden del habla. En la Figura 5 puede ser visto otro ejemplo de clasificación de desórdenes del habla [39].

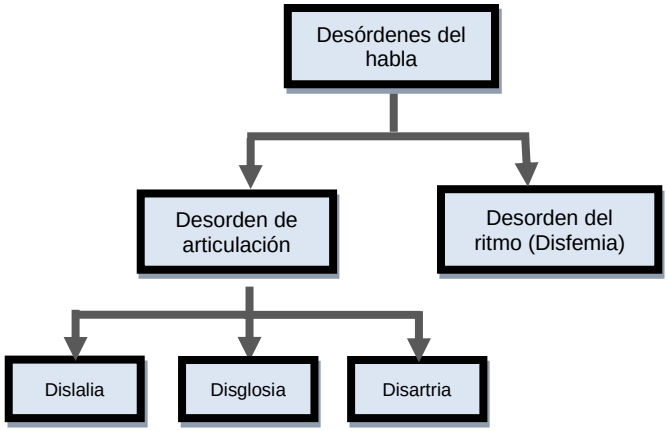

Fig. 5 Clasificación adicional para desórdenes del habla.

Al estar la ontología enfocada en el diagnóstico y terapia de los desórdenes fonéticos y de fluidez los términos relevantes a estos serán empleados. **Cabe aclarar en este punto que debido a que la mayoría de los documentos encontrados en la web relevantes al dominio de los desórdenes del habla se encuentran en idioma inglés, se hizo la búsqueda de los términos relevantes en ambos idiomas (inglés y español) para obtener un corpus más completo**. La Tabla 1 muestra la primera versión del diccionario primario.

Tabla 1. Lista de términos del diccionario primario

No.	Término(s)	Término(s)
1	Speech	Habla
2	Disorder	Desorden
3	Dyslalia	Dislalia
4	Dysglosia	Disglosia
5	Dysarthria	Disartria
6	Dysphemia	Disfemia
7	Speech sound disorder	Desorden del habla
8	Childhood-onset fluency disorder	Desorden de fluidez en la infancia
9	Communication disorder	Desorden de la comunicación
10	Articulation disorder	Desorden de articulación
11	Rhythm disorder	Desorden del ritmo
12	Therapy	Terapia
13	Speech therapy	Terapia del habla
14	Logopedic therapy	Terapia logopédica
15	Speech development	Desarrollo del habla

Comenzando con los términos obtenidos directamente de las ramas de la taxonomía que está relacionada al dominio, términos como *desorden de la comunicación* / *communication disorder*, *desorden del habla / speech sound disorder* o *desorden de la fluidez en la infancia / childhood-onset fluency disorder* se incluyen en el diccionario. Como la ontología contendrá información sobre terapias es también deseable que se incluyan términos como *terapia del habla / speech therapy, terapia logopédica / logopedic therapy* y *desarrollo del habla / speech development.*

El siguiente paso es el uso de este diccionario para recolectar los documentos del corpus para la ontología utilizando un web crawler escrito en lenguaje Python con la ayuda de librerías *HTMLparser, urlopen* y *BeautifulSoup* [40]. Usando algunas páginas relevantes al dominio de los desórdenes del habla (*asha.org*, *medlineplus.com*, *Google Scholar*, *NCBI* -National Center for Biotechnology Information, *NICHCY*- National Dissemination Center for Children with Disabilities , *NIDCD*- National Institute on Deafness and Other Communication Disorders, etc) el recorrido de tales sitios es iniciado para buscar cada término del diccionario uno a la vez y recuperar información en cada sitio visitado, almacenar tal información en un archivo y posteriormente guardar los links para visitar páginas internas dentro de la *página semilla* proporcionada al crawler como parámetro. Un parámetro adicional al

crawler puede ser el número máximo de páginas a visitar en la búsqueda del término. Después de recuperar información relevante para todos los términos del diccionario la primera versión del corpus está terminada, pero el procesamiento del corpus no ha terminado.

3.2.2 Pre-procesamiento de la información

El siguiente paso referente al pre-procesamiento de la información es hecho a través de varios algoritmos que normalizan el texto contenido en el corpus. Algoritmos para remover caracteres de escape, caracteres Unicode, signos de puntuación, stop words, convertir a texto plano y remover mayúsculas son muy útiles para la limpieza de la información antes de ser analizada [41][42]. Nuevamente, con algunas rutinas en lenguaje Python son ejecutados los algoritmos para limpiar la información. Una vez que toda la información recopilada en el corpus es normalizada se procede al siguiente paso.

En este paso, algoritmos de recuperación de información son implementados. Algoritmos como *lematización* y *Word frequency* son utilizados [42]. Después de este paso una nueva lista de términos para el diccionario extendido es obtenida. Los términos más frecuentes encontrados en el corpus son tomados y comparados contra los términos del diccionario primario. En la siguiente sección es presentada tal comparación.

3.2.3 Diccionario extendido

Para recuperar los documentos de un corpus, uno de los métodos más utilizados es comprobar la presencia o ausencia de las palabras que forman la consulta en cada documento, lo cual implicaría que la recuperación se mide en términos absolutos. Pero este método de recuperación no es aplicable a los modelos booleano, vectorial y probabilístico en los que se basa la recuperación de información moderna. Por otro lado, también es posible considerar una frecuencia de aparición de los términos mayor, para determinar un documento como más idóneo para resolver una consulta o búsqueda.

Después de aplicar al corpus el pre-procesamiento descrito en la sección anterior y los algoritmos de recuperación de información, los términos mostrados en la Tabla 2 fueron algunos de los más frecuentes dentro del corpus y en frecuencia de documentos *DF*. Al tener mayor cantidad de documentos en idioma inglés conformando el corpus, los términos frecuentes también tuvieron resultados en ese idioma.

Tabla 2. 15 términos más frecuentes en el corpus.

No.	Término	Frecuencia en el corpus	Número de documentos con el término (DF)
1	Speech	13,032	980
2	Child	9,225	604
3	Language	8,250	993
4	Disorder	6,338	952
5	Health	4,940	739
6	Therapy	3,757	788
7	Help	3,738	902
8	Word	3,720	458
9	Information	3,547	469
10	Sound	3,396	412

11	Development	2,916	470
12	Research	2,720	307
13	Service	2,636	205
14	Communication	2,243	561
15	Medical	1,813	476

Pero en todo caso, lo idóneo es estimar el valor de cada término en cuanto a la recuperación, representación y discriminación de los contenidos en el corpus documental, ya que son muchos varios factores determinantes. La *ponderación* de los términos es un proceso que tiene como finalidad conocer la importancia de los términos en un documento y así realizar su posterior recuperación. Esto implica que se debe determinar la capacidad de los términos para representar el contenido de los documentos en el corpus, que permiten identificar cuáles son relevantes para el dominio del corpus. Al valor que es capaz de determinar esto se le llama "peso del término" o también "ponderación del término" y para calcularlo se debe determinar la "Frecuencia de aparición del término" o TF, y a su vez la "Frecuencia inversa del documento para un término" o IDF.

Factor TF: Term Frequency

El factor TF es la suma de todas las ocurrencias o el número de veces que aparece un término (*t*) en un documento (*d*). A este tipo de frecuencia de aparición también se la denomina "Frecuencia de aparición relativa" por que atañe a un documento en concreto y no a toda la colección. Dicho de otra forma, el número de veces que este se repite en el documento, lo que permite determinar su capacidad de representación. Su finalidad es puramente representativa y comprende los siguientes casos:

- Frecuencia de aparición TF baja. Representatividad elevada.
- Frecuencia de aparición TF media.
- Frecuencia de aparición TF alta. Muy baja representatividad.

Y se expresa con la siguiente fórmula

$$\mathrm{tf}(t, d) = f_{t,\,d.}$$

Factor IDF: Inverse Document Frequency

El factor IDF de un término es inversamente proporcional al número de documentos en los que aparece dicho término. Esto significa que cuanto menor sea la cantidad de documentos, así como la frecuencia absoluta de aparición del término, mayor será su factor IDF y a la inversa, cuanto mayor sea la frecuencia absoluta relativa a una alta presencia en todos los documentos de la colección, menor será su factor discriminatorio. Es el coeficiente que determina la capacidad discriminatoria del término de un documento con respecto a la colección. Es decir, distinguir la homogeneidad o heterogeneidad del documento a través de sus términos. Su finalidad es discriminatoria y presenta los siguientes casos:

- Poder discriminatorio bajo. El término es genérico y aparece en la mayoría de los documentos.
- Poder discriminatorio medio.
- Poder discriminatorio alto. El término es especializado y aparece en pocos documentos.

La fórmula para calcularlo es la siguiente:

$$idf(t, D) = log \frac{N}{|\{d \in D: t \in d\}|}$$

con

- N número total de documentos en el corpus $N = |D|$
- $|\{d \in D: t \in d\}|$ número de documentos donde el término t aparece, es decir, $tf(t, d) \neq 0$ mencionada anteriormente como *DF*

si el término no está en el corpus, esto genera una división por cero. Por lo tanto, es común ajustar el denominador sumándole 1.

Ponderación TF-IDF

Para calcular el peso de un término en un documento se toma el producto de su frecuencia de aparición en dicho documento (*TF*) y su frecuencia inversa de documento (*IDF*) como lo muestra la siguiente fórmula en la Figura 6.

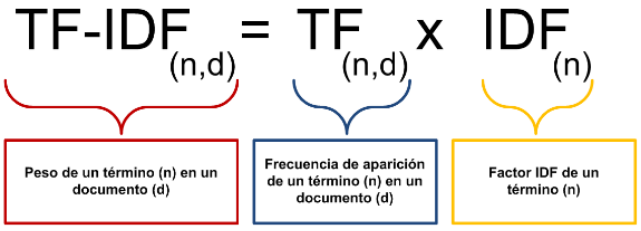

Fig 6. Fórmula de la ponderación TF-IDF

Al aplicar la ponderación *IDF* a los términos del corpus encontramos que términos relevantes para el dominio son representados con su valor discriminatorio como se muestra en la Tabla 3. En tal caso, se calculó el IDF correspondiente a 20 términos presentes en una colección de 1097 documentos. *DF* representa la frecuencia de documentos o lo que es lo mismo, el número de documentos en los que el término aparece. Finalmente se aplica la fórmula *IDF*, en la que se observa la cualidad potencial inversamente proporcional de la capacidad discriminatoria. Al ordenar de mayor a menor los términos según su coeficiente *IDF* se observa que los que mayor poder discriminatorio son los que menos valor *DF* tienen.

Tabla 3. Ponderación de términos relevantes en el corpus.

No.	Término	Frecuencia en el corpus	DF	IDF
1	Dysglosia	4	2	3.73
2	Audiometry	14	5	3.34
3	Logopedic	16	11	2.99
4	Aphonia	30	13	2.92
5	Dysphemia	46	28	2.59
6	Phonation	62	42	2.41
7	Dyslalia	89	55	2.29
8	Dysphonia	91	81	2.13
9	Phonetic	155	96	2.05
10	Dyspraxia	105	100	2.04
11	Pronunciation	178	113	1.98
12	Dysphagia	136	125	1.94
13	Lisp	230	189	1.76
14	Dysarthria	475	278	1.59
15	Fluency	444	305	1.55
16	Stammering	498	346	1.50
17	Phonological	871	357	1.48
18	Stuttering	1059	363	1.48
19	Apraxia	647	370	1.47
20	Therapist	685	415	1.42

Observando los datos obtenidos con la frecuencia de términos y ponderación, no todos los términos propuestos en el diccionario primario son igualmente relevantes dentro de los documentos del corpus. Utilizando estos términos basados en su peso en cada documento del corpus, el *web crawler* puede ser alimentado nuevamente con los términos más recientes obtenidos del corpus y así, recolectar más documentos relevantes.

Sinónimos, hipónimos e hiperónimos

Otra manera de complementar el corpus es incluir sinónimos de los términos propuestos originalmente para recuperar más documentos [43]. Una cantidad significativa de sinónimos fue encontrada para ser incluida en tal lista, algunos de ellos están listados en la Tabla 4.

Tabla 4. Sinónimos para algunos de los términos propuestos.

No.	Término(s) propuestos	Sinónimos
1	Speech	Conversation, locution, expression, language, articulation.
2	Disorder	Irregularity, impairment, deficit.
3	Dyslalia	Dysphasia.
4	Dysarthria	Aphasia.
5	Dysphemia, Childhood-onset fluency disorder, Rhythm disorder.	Stammering, stuttering.
6	Speech sound disorder, Communication disorder, Articulation disorder.	Speech impairment, speech impediment, speech defect, delayed speech, speech deficit, speech deficiency, speech disturb-

		ance, misarticulation, phonological disorder, phonological delay, phonological impairment, verbal disorder.
7	Therapy.	Treatment, Care.
8	Speech therapy, Logopedic therapy.	Language therapy, Articulation therapy, Speech treatment.
9	Speech development	Speech progress, Speech improvement, Speech maturation, Speech progression.

Al aplicar nuevamente los pasos de *crawling*, pre-procesamiento y algoritmos de IR más documentos son añadidos al corpus y una nueva lista de los términos más frecuentes es obtenida.

La mayoría de los términos más frecuentes y con valor *DF* alto, obtenidos después de esta expansión en el diccionario resultaron ser los mismos que los obtenidos en el paso previo al uso de sinónimos, sólo variando el orden de aparición en la lista. En los casos de términos como *child* y *language* resultaron más frecuentes cuando fueron utilizados sinónimos como semillas.

Una manera más de expandir nuestro corpus y el diccionario de términos es utilizando las relaciones hiponímia/hiperonímia entre términos ya que una de las maneras de organizar nuestro vocabulario consiste en la formación de campos semánticos. Generando con ayuda de WordNet una lista de algunos de los términos frecuentes con sus hipónimos e hiperónimos podemos nuevamente alimentar el crawler para recuperar más documentos [3]. La Tabla 5 muestra algunos de los términos frecuentes con sus hipónimos y/o hiperónimos.

Tabla 5. Listado de algunos términos frecuentes con sus respectivos hipónimos/hiperónimos.

Hiperónimo	Término	Hipónimo
Auditory communication	Speech, Oral communication	Pronunciation, locution, words
Physical condition	Disorder	Speech defect, functional disorder, organic disorder, ailment, anarthria, aphonia, lallation, lisp, lambdacism
Defect of speech	Dysarthria	----
Medical aid, medical care	Therapy	Speech therapy
Auditory communication	Language	Spoken communication, Word
Wellbeing	Health	----
Languaje unit	Word	----

Se ejecutan nuevamente los pasos realizados en la extensión anterior del diccionario (*crawling*, pre-procesamiento y algoritmos de IR) y más documentos son añadidos al corpus y se obtienen nuevamente los términos con más peso para los documentos dentro del corpus.

N-gramas

En las etapas anteriores de expansión del diccionario de términos, tanto los términos propuestos para el diccionario primario como los términos encontrados a través de sinónimos, hipónimos e hiperónimos se tratan no sólo de términos formados por una sola palabra sino por colocaciones o n-gramas. En el caso del dominio de desórdenes del habla los n-gramas

más representativos resultaron ser los trigramas, la Tabla 6 muestra algunos de los trigramas más frecuentes y con relevancia para el dominio dentro del corpus.

Tabla 6. Trigramas frecuentes y relevantes encontrados en el corpus

No.	Término	Frecuencia en el corpus (TF)
1	Speech articulation disorder	18
2	Speech hearing disorder	24
3	Child language development	24
4	Speech language impairment	25
5	Child speech delay	27
6	Speech language therapy	36
7	Childhood-onset fluency disorder	39
8	Stuttering fluency children	41
9	Speech language development	45
10	Language hearing research	60
11	Language communication disorder	63
12	Speech language pathologist	67
13	Disorder developmental verbal	82
14	Childhood apraxia speech	88
15	Speech sound disorder	95

Estos trigramas también serán tomados como términos claves para la construcción de la ontología junto con los demás términos extraídos en los pasos anteriores.

Una vez que se tiene consistencia en los documentos del corpus al analizar los términos más comunes en diferentes etapas se puede proceder a un siguiente paso para la extracción de conceptos y aplicar algoritmos como el uso de patrones para la identificación de más relaciones semánticas.

A continuación, algunos datos descriptivos del corpus después de los pasos anteriores son mostrados en la Tabla 7.

Tabla 7. Datos descriptivos del corpus

Número de términos iniciales utilizados para la recuperación de documentos.	**15**
Número de documentos añadidos manualmente al corpus	**50**
Número de documentos obtenidos al final del proceso de crawling	**1097**
Tokens (palabras) en el corpus	**1,214,287**
Types (tipos) en el corpus	**74,664**
Tamaño en bytes del corpus en texto plano	**8,416,391**

3.3 Exactitud en las búsquedas: Precision and Recall

La exactitud en las búsquedas es medida a menudo utilizando las métricas de recuperación de información *Precision* y *Recall*. Siendo estas medidas buenas para llevar un registro de la efectividad de una búsqueda en particular.

Precision mide el número de documentos realmente relacionados en el conjunto recuperado de documentos obtenidos. *Recall* mide el número de documentos relacionados recuperados comparados al número total de documentos relacionados en el corpus. Dicho de otra manera, *Precision* es la fracción de la información recuperada que es relevante y *Recall* es la fracción de la información relevante que se recupera. Estos dos índices son utilizados de manera extensa para caracterizar la efectividad en los sistemas de recuperación de información. Podemos expresarlos mediante las siguientes fórmulas:

$$Precision = \frac{\{\text{Número de documentos relevantes recuperados}\}}{\{\text{Total documentos recuperados}\}}$$

$$Recall = \frac{\{\text{Número de documentos relevantes recuperados}\}}{\{Total \text{ documentos relevantes}\}}$$

Un número alto como resultado de Precision implica que un alto porcentaje de documentos relevantes son recuperados y sólo un pequeño número de documentos no relevantes fueron categorizados como relevantes. Un número alto de Precision es de ayuda al establecer una segunda revisión con intervención humana.

Un número alto de Recall sugiere que el sistema de recuperación de información es efectivo al recuperar un alto porcentaje de documentos relevantes, y pocos documentos relevantes son dejados en la colección sin recuperar. Un Recall muy pequeño sugiere mejorar los métodos de recuperación automática.

Hablemos ahora de la Matriz de Confusión, para avanzar con los conceptos de Precision y Recall. La Matriz de Confusión es muy útil al calcular Precision y Recall. Una matriz de confusión binaria muestra cuatro diferentes salidas. Los valores predichos forman las columnas y los valores reales u observados forman los renglones como se muestra en la Figura 7.

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Fig. 7 Matriz de Confusión

Así, tenemos 4 diferentes salidas:

- Verdaderos positivos (predicho 1, real también 1)
- Falsos positivos (predicho 1, real 0)
- Falsos negativos (predicho 0, real 1)

- Verdaderos negativos (predicho 0, real también 0)

Un factor que también debe ser tomado en consideración es el número total de documentos. Cuanto más grande un corpus los valores de Precision y Recall caen considerablemente. La intersección de Precision y Recall es un punto crítico de optimización.

Para solucionar el problema anterior de optimización existe otra medida que suele ser útil, la conocida como F-Measure o F_1 Score, la cual combina ambas medidas en un solo valor.

$$F_1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

Esta medida por si sola es representativa de efectividad en las búsquedas.

En este caso a manera de validación utilizando una parte de los documentos del corpus se busca medir los resultados obtenidos al hacer búsquedas dentro de él por medio de las medidas Precision y Recall. Utilizando el 10% de los documentos del corpus que fueron previamente etiquetados en 3 categorías de clasificación de desórdenes del habla se procedió a hacer búsquedas dentro del corpus y calcular ambas medidas para ver la efectividad de la recuperación y la relevancia de los documentos recuperados para proceder al próximo paso de extracción de conceptos y relaciones entre ellos para la construcción de la ontología. En la siguiente matriz de confusión de la Tabla 8 pueden verse los valores de las diferentes pruebas con los términos en cada búsqueda.

Tabla 8. Matriz de confusión del corpus

Documentos en el conjunto de prueba	**Desorden de articulación**	**Desorden de ritmo y fluencia**	**Desorden de voz y resonancia**
Desorden de articulación	70%	18%	12%
Desorden de ritmo y fluencia	19%	68%	13%
Desorden de voz y resonancia	3%	15%	82%

En este caso los valores de Precision, Recall y F1 score para cada búsqueda son mostrados a continuación en la Tabla 9.

Capítulo 4

Desarrollo de la ontología

Primero, el escenario en el cuál es aplicable la ontología de ser definido, seguido por la generación de las llamadas “preguntas de competencia” en lenguaje natural, cuyo objetivo es determinar el alcance de la ontología. Estas preguntas y sus correspondientes respuestas son utilizadas para extraer los principales conceptos, así como sus relaciones, propiedades y axiomas dentro de la ontología. La formalidad de este método nos permite transformar escenarios informales en modelos computacionales. Los elementos para el diseño son los siguientes: construcción de la taxonomía para la base de conocimiento a ser representada; atributos y relaciones entre clases; y reglas y atributos de las instancias.

El uso de ontologías para representar una base de conocimiento dentro de cierto dominio tiene el propósito de facilitar en entendimiento de tal dominio y obtener mejor información sobre el tema. La información relevante sobre los desórdenes del habla es la clasificación de cada desorden con su propia subclasificación para correctamente catalogar del desorden presentado por cada paciente, los síntomas y signos producidos por cada desorden -esas son las pistas que el terapeuta debería buscar-, la etiología que podría afectar el curso y resultado de la terapia y las diferentes partes de la terapia, primero con una estrategia de evaluación y después con una estrategia de intervención guiada por el terapeuta. Una vez que el escenario para el área de competencia de la ontología es definido, el conjunto de taxonomías puede ser utilizado para llegar a una definición de las clases de la ontología como se muestran en la Figura 8, y las relaciones entre ellas; una serie de preguntas esperando a ser respondidas a través de consultas a la ontología también será definida. Una definición formal es hecha para las clases y sus atributos, así como para la descripción de las relaciones y axiomas de la ontología.

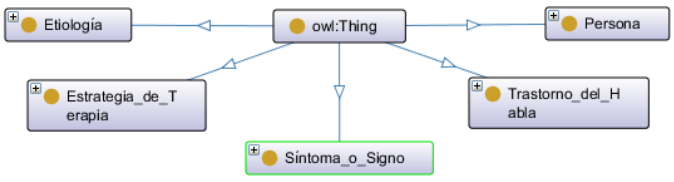

Fig. 8. Clases principales de la ontología.

4.1 Preguntas de Competencia

Las preguntas de competencia son una parte importante en los pasos de diseño de la ontología porque nos permiten definir el dominio y alcance de la ontología.

La ontología propuesta busca respuesta a preguntas como las siguientes:

- ¿Cuál es el desorden del habla más común en niños?
- ¿Cuáles son los síntomas de un desorden del habla particular?
- ¿Cuántos tipos de desórdenes del habla existen?
- ¿Cuál es la causa de cierto tipo de desorden?
- ¿Cuál es la terapia para un desorden del habla?
- ¿Qué es la dislalia?
- ¿A qué edad puede ser notado un desorden del habla?

La base de conocimiento de la ontología debe ser capaz de responder tales preguntas, en esta fase las preguntas son presentadas en lenguaje natural.

4.2 Definición de clases

Las siguientes entidades son algunas de las encontradas después de un análisis del escenario del área de competencia. Una estrategia mixta fue utilizada (top-down y bottom-up) para identificar los conceptos principales [24] mostrados en la tabla 10 .

Tabla 10. Definición de las clases.

Clase	Definición
Trastorno_del_habla	Esta clase contiene la taxonomía complete de desórdenes del habla.
Etiología	Esta clase incluye la taxonomía de las diferentes causas de los desórdenes del habla.
Persona	Esta clase incluye a los diferentes individuos que son diagnosticados con (*Paciente*) o diagnostican un desorden del habla (*Terapeuta*).
Síntoma_o_Signo	Clase que incluye los diferentes síntomas o signos que son presentados por un paciente con un desorden del habla.
Estrategia_de_Terapia	Clase que contiene las dos partes principales de las acciones de terapia aplicada al *Paciente*.
Trastorno_de_la_articulación	Subclase de *Trastorno_del_Habla*, consistente en la dificultad de pronunciar sonidos.
Trastorno_del_ritmo_y_fluidez	Subclase de *Trastorno_del_Habla*, se refiere a una alteración en el ritmo del habla.
Trastorno_de_la_voz_y_resonacia	Subclase de *Trastorno_del_Habla*, en una alteración de la voz en el volumen, tono o timbre.

Las clases y subclases previas pueden ser vistas en el siguiente diagrama jerárquico de la Figura 9 generado con el software Protégé [44].

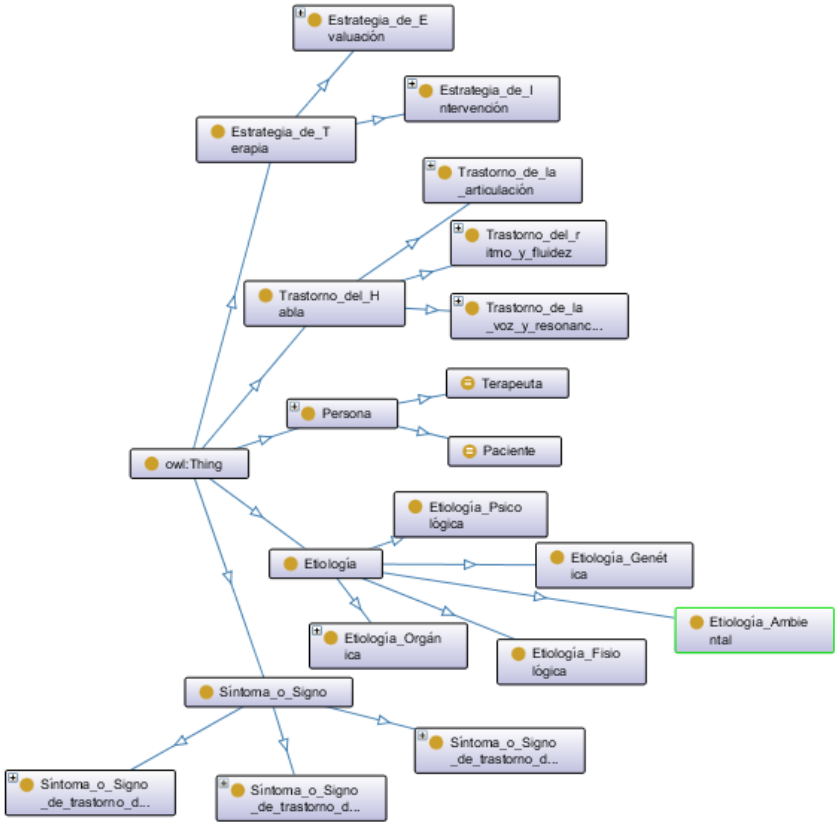

Fig. 9. Diagrama de jerarquía de clases.

4.3 Descripción de relaciones

Dentro de la ontología existen ciertas limitantes con respecto a las clases de la ontología misma. Para comenzar a describir estas limitantes es necesario considerar las relaciones entre clases. En la Tabla 11 algunas de las relaciones identificadas son explicadas, puede notarse que cada una tiene una relación inversa también representada dentro de la ontología.

Table 11. Descripción de relaciones entre clases.

Relación	Dominio	Rango	Inversa	Cardinalidad
Afecta_a	Trastorno_del_habla	Paciente	Padece_un	N:1
Aplica_una	Terapeuta	Estrategia_de_Terapia	Es_aplicada_por	1:N
Evalúa_un	Estrategia_de_Evaluación	Trastorno_del_Habla	Es_evaluado_por	1:1
Da_terapia_a	Terapeuta	Paciente	Recive_terapia_de	1:N
Tiene_causa	Trastorno_del_habla	Etiología	Es_causa_de	1:N
Interviene_un	Estrategia_de_Intervención	Trastorno_del_Habla	Es_intervenido_por	1:1
Es_manifestación_de	Síntoma_o_Signo	Trastorno_del_Habla	Se_manifiesta_con_un	N:1
Es_presentado_por	Síntoma_o_Signo	Paciente	Manifiesta_un	N:1

Estas relaciones pueden ser observadas de manera visual en el siguiente diagrama. Las relaciones pueden ser representadas como un grafo donde cada clase es presentada como un nodo y las aristas entre nodos son las relaciones entre clases (Ver Fig. 5).

Los axiomas que definen las reglas de la ontología son establecidos por las características y las restricciones existenciales de las relaciones no taxonómicas. Los *Object properties* en Protégé son las relaciones entre las clases.

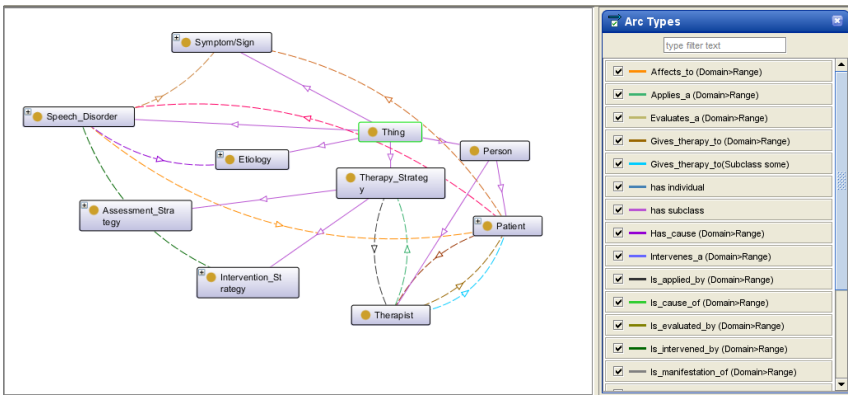

Fig. 1. Diagrama de relaciones entre clases.

Las características de la relación pueden ser vistas como funciones y en Protégé son llamadas *property characteristics*; estas características que pueden ser asociadas a una propiedad (relación) pueden ser *Funcionales*, *Funcionales Inversas*, *Transitivas*, *Simétricas*, *Antisimétricas*, *Reflexivas* e *Irreflexivas*. Algunas de estas características son asignadas a cada uno de los *object properties* dependiendo del tipo de relación entre clases que sea representado por ellos; un ejemplo puede ser visto en la Figura 6 donde algunas características de la relación (*property characteristics*) son asignadas a la relación (*object property*) *Evalúa_un* y a su inversa *Es_evaluado_por*, siendo *Funcional*, *Funcional Inversa* –en el caso de la propiedad inversa–, *Asimétrica* e *Irreflexiva* las características asignadas dependiendo del comportamiento de cada objeto.

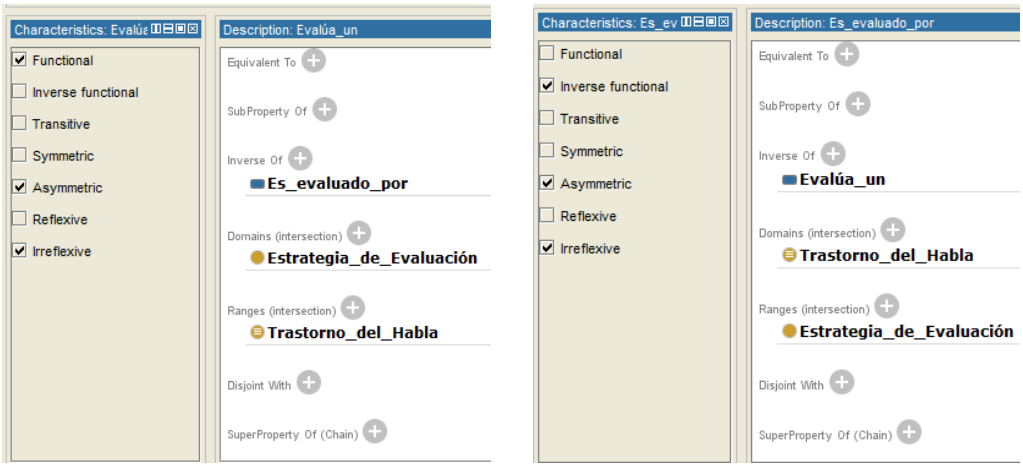

Fig. 2. Características de las relaciones.

La característica *Funcional* indica para una relación dada, que puede haber *a lo más* una clase rango que es relacionada a la clase dominio por medio de la propiedad. Y si la propiedad es

Funcional Inversa entonces significa que la propiedad inversa es funcional. Si una propiedad P es *Asimétrica*, y la propiedad relaciona a la clase *a* con la clase *b*, entonces la clase *b* no puede estar relacionada a la clase *a* por medio de la propiedad P. Y finalmente, si la propiedad P es *Irreflexiva*, puede ser descrita como una propiedad que relaciona la clase *a* con la clase *b*, donde la clase *a* y la clase *b* no son la misma. En la siguiente imagen se muestra al grafo que representa esas relaciones (Ver Fig. 7).

Otras restricciones que ayudan a describir y definir las clases son las restricciones cuantificables, en este caso las restricciones existenciales y universales. Principalmente las restricciones cuantificables encontradas en las ontologías son existenciales; esto significa una clase de individuos que tiene *al menos una* (alguna) relación con una propiedad específica a un individuo que es miembro de una clase específica.

Fig. 3. Las relaciones *Evalúa_un_a* y *Es_evaluado_por* representadas como grafo.

4.4 Prueba de consistencia

Con el fin de probar la consistencia de la ontología construida, el razonador lógico de Protégé es utilizado con la técnica *probe class*. Esto significa agregar una clase inconsistente para probar la integridad de la ontología. En este caso, se agregó una nueva clase: *TrastornoInconsistente*, que es una subclase de *Trastorno_de_la_articulación* y *Trastorno_de_la_voz_y_resonancia*, simultáneamente. Después de invocar al razonador para probar la consistencia de la clase agregada, se muestra un error porque las súper clases a las que pertenece son disjuntas entre sí. La prueba de consistencia se muestra en las Figuras 8 a 11, junto con las propiedades de clase definidas y el error resultante.

Bibliografía

[1] B. (Ohio S. U. Chandrasekaran, J. R. (Ohio S. U. Josephson, and V. R. (University of A. Benjamins, "What Are Ontologies and Why Do We Need Them?," *IEEE Intell. Syst. their Appl.*, vol. 14, no. 1, pp. 20–26, 1999.

[2] Cycorp Company USA, "CYC: Logical Reasoning with the World's Largest Knowledge Base," 2017. [Online]. Available: http://www.cyc.com/. [Accessed: 27-Aug-2018].

[3] Princeton University, "About WordNet - WordNet." [Online]. Available: http://wordnet.princeton.edu/wordnet/. [Accessed: 17-Nov-2017].

[4] NICHCY, "Trastornos del habla o lenguaje," vol. 0285, pp. 1–4, 2010.

[5] "Disability in Mexico | Global Disability RightsNow!" [Online]. Available: http://www.globaldisabilityrightsnow.org/infographics/disability-mexico. [Accessed: 27-Jun-2017].

[6] WHO (World Health Organization), "World report on disability 2011," *Am. J. Phys. Med. Rehabil. Assoc. Acad. Physiatr.*, 2011.

[7] DOF, "PROGRAMA NACIONAL PARA EL DESARROLLO Y LA INCLUSIÓN DE LAS PERSONAS CON DISCAPACIDAD 2014?2018," *DOF 30/04/2014*, vol. 17852, no. 10, pp. 1–17, 2014.

[8] K. Loudon, *Developing Large Web Applications*, First Edit. California: O'Reilly Media, 2010.

[9] Robin, "What is Corpus?," *Natural Language Processing*, 2009. [Online]. Available: http://language.worldofcomputing.net/linguistics/introduction/what-is-corpus.html. [Accessed: 20-Aug-2017].

[10] R. Lozano-Rubí, X. Pastor, and E. Lozano, "OWLing clinical data repositories with the ontology web language," *J. Med. Internet Res.*, vol. 16, no. 8, p. e14, 2014.

[11] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*. 2013.

[12] A. Bashir, "Language Intervention and the Curriculum," *Semin. Speech Lang.*, vol. 10, no. 03, pp. 181–191, Aug. 1989.

[13] J. R. Gallardo Ruíz, J. L. Gallego Ortega, J. Valcárcel Castilla, C. Pardal Rivas, and V. Vilchez García, *Manual de logopedia escolar: un enfoque práctico*. Málaga: Aljibe, 2000.

[14] M. Coll-Florit, J. M. Vila-Rovira, G. Aguado, A. Fernández-Zúñiga, S. Gambra, and E. Perelló, *Trastornos del habla y de la voz*, 1a ed. Barcelona: Editorial UOC, 2014.

[15] E. A. Marecos, "El diagnóstico diferencial," *Rev. Posgrado en la Vía Cátedra Med.*, vol. 128, 2003.

[16] "American Speech-Language-Hearing Association | ASHA." [Online]. Available: http://www.asha.org/. [Accessed: 05-Jul-2017].

[17] T. I. M. Berners-lee, J. Hendler, and O. R. A. Lassila, "The Semantic Web," *Sci. Am.*, no. May, pp. 1–4, 2001.

[18] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowl. Acquis.*, vol. 5, no. 2, pp. 199–220, 1993.

[19] A. Lozano Tello, "Ontologías en la Web Semántica," *I Jornadas Ing. Web'01*, vol. 1, pp. 1–4, 2001.
[20] E. Ramos, H. Núñez, and R. Casañas, "Schemes to evaluate singular ontologies for a knowledge domain," *Rev. Venez. Inf. Tecnol. y Cocnocimiento*, vol. 6, no. 1, pp. 57–71, 2009.
[21] L. Feigenbaum, I. Herman, T. Hongsermeier, E. Neumann, and S. S., "The Semantic Web in Action," *Sci. Am.*, vol. 297, no. December, pp. 90–97, 2007.
[22] O. Corcho, M. Fernández-López, and A. Gómez-Pérez, "Methodologies, tools and languages for building ontologies. Where is their meeting point?," *Data Knowl. Eng.*, vol. 46, no. 1, pp. 41–64, 2003.
[23] M. Gruninger and M. S. Fox, "Methodology for the Design and Evaluation of Ontologies," in *Workshop on Basic Ontological Issues in Knowledge Sharing*, 1995, no. IJCAI-95.
[24] A. Gómez-Pérez, M. Fernández-López, O. Corcho, and A. Gomez-Perez, *Ontological Engineering*, 1st ed. Springer, 2004.
[25] S. Sharma, E. C. Ward, C. Burns, D. Theodoros, and T. Russell, "Assessing dysphagia via telerehabilitation: Patient perceptions and satisfaction," *Int. J. Speech. Lang. Pathol.*, vol. 15, no. 2, pp. 176–183, 2013.
[26] O. A. Schipor, S. G. Pentiuc, and M. D. Schipor, "Automatic assessment of pronunciation quality of children within assisted speech therapy," *Elektron. ir Elektrotechnika*, vol. 122, no. 6, pp. 15–18, 2012.
[27] M. L. Martín Ruiz, M. Á. Valero Duboy, C. Torcal Loriente, and I. Pau de la Cruz, "Evaluating a web-based clinical decision support system for language disorders screening in a nursery school.," *J. Med. Internet Res.*, vol. 16, no. 5, p. e139, May 2014.
[28] V. Robles-Bykbaev, M. L??pez-Nores, J. Pazos-Arias, D. Quisi-Peralta, and J. Garc??a-Duque, "An Ecosystem of Intelligent ICT Tools for Speech-Language Therapy Based on a Formal Knowledge Model," *Stud. Health Technol. Inform.*, vol. 216, pp. 50–54, 2015.
[29] F. Chuchuca-Mendez, V. Robles-Bykbaev, P. Vanegas-Peralta, J. Lucero-Saldana, M. Lopez-Nores, and J. Pazos-Arias, "An educative environment based on ontologies and e-learning for training on design of speech-language therapy plans for children with disabilities and communication disorders," *CACIDI 2016 - Congr. Aergentino Ciencias la Inform. y Desarro. Investig.*, 2016.
[30] R. Ilda, B. Torres, I. V. López, J. Luis, and D. Suarez, "Aplicación Móvil para la Adquisición de Lenguaje En Niños Con Trastorno De Habla," no. 122, pp. 40–56, 2016.
[31] D. Kalra, T. Beale, and S. Heard, "The OpenEHR foundation." [Online]. Available: http://www.openehr.org/home. [Accessed: 02-Jul-2017].
[32] V. Robles-Bykbaev, M. López-Nores, J. Pazos-Arias, J. García-Duque, and J. Ochoa-Zambrano, "Modelling domain knowledge of speech and language therapy with an OWL ontology and OpenEHR archetypes," *Heal. 2015 - 8th Int. Conf. Heal. Informatics, Proceedings; Part 8th Int. Jt. Conf. Biomed. Eng. Syst. Technol. BIOSTEC 2015*, pp. 585–591, 2015.
[33] H. M. Haav, "A Semi-automatic Method to Ontology Design by Using FCA," *CLA*,

pp. 13–24, 2004.
[34] F. Braga and N. Ebecken, “A semi-automatic method for extracting a taxonomy for nuclear knowledge using hierarchical document clustering based on concept sets Fabiane Braga,” *Int. J. Nucl. Knowl. Manag.*, vol. 6, no. 2, pp. 155–169, 2013.
[35] A. Maedche and S. Staab, “Semi-automatic engineering of ontologies from text,” *Proc. 12th Int. Conf. Softw. Eng. Knowl. Eng.*, pp. 231–239, 2000.
[36] M. Wynne, “Developing Linguistic a Guide to Good Practice Corpora :,” 2005.
[37] D. Evans, “Corpus building and investigation for the Humanities : An on-line information pack about corpus investigation techniques for the Humanities,” *Linguistics*, pp. 15–16, 2004.
[38] R. Mitchell, *Web scraping with Python: collecting data from the modern web.*, First Edit. O’Reilly Media, Inc., 2015.
[39] M. del C. Busto Barcos and M. P. Martínez Guijarro, *Manual de logopedia escolar: niños con alteraciones del lenguaje oral en educación infantil y primaria.* CEPE. CIENCIAS DE LA EDUCACION PREESCOLAR Y ESPECIAL., 2007.
[40] Python Language Reference, “Python Software Foundation,” *Python Language Reference, version 3.6.1*, 2010. [Online]. Available: https://www.python.org/. [Accessed: 09-Sep-2017].
[41] Caio Miyashiro, “Text Mining and Natural Language Processing - Preprocessing,” 2015. [Online]. Available: http://rstudio-pubs-static.s3.amazonaws.com/67435_ca0769f0dbbb4fc4bda5e4535e21fb54.html. [Accessed: 09-Sep-2017].
[42] X. Zhu, “Common Preprocessing Steps,” *CS769 Spring 2010 Adv. Nat. Lang. Process.*, pp. 1–3, 2010.
[43] S. D. Knapp, *The contemporary thesaurus of search terms and synonyms: A guide for natural language computer searching*, 2nd ed. Greenwood Publishing Group, 2000.
[44] M. A. Musen, “The Protégé project: A look back and a look forward.,” *AI Matters. Assoc. Comput. Mach. Specif. Interes. Gr. Artif. Intell.*, vol. 1, no. 4, 2015.
[45] U. Prot *et al.*, “A Practical Guide To Building OWL Ontologies Using Protégè 4 and CO-ODE Tools Edition 1.3,” *Matrix*, pp. 0–107, 2011.