VNIVERSIDAD NACIONAL
AVFNOMA DE
MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE FILOSOFÍA Y LETRAS

COLEGIO DE LETRAS HISPÁNICAS

DETECCIÓN DE SIMILITUD TEXTUAL MEDIANTE CRITERIOS DE DISCURSO Y SEMÁNTICA

TESIS

QUE, PARA OBTENER EL TÍTULO DE
LICENCIADA EN LENGUA Y LITERATURAS HISPÁNICAS,
PRESENTA

BRENDA GABRIELA CASTRO ROLÓN

ASESOR: DR. GERARDO EUGENIO SIERRA MARTÍNEZ
CO-ASESOR: DR. JUAN MANUEL TORRES-MORENO

CIUDAD UNIVERSITARIA, 2011

Agradecimientos

Agradezco a todas las personas involucradas en el desarrollo de esta tesis. Quiero resaltar que a pesar de que es mi nombre el que aparece en la portada, es un trabajo de grupo.

En primer lugar, agradezco a mis padres, sin ellos ni yo ni este trabajo hubiéramos sido posibles. Les debo todo. A mi madre, que es un ejemplo de esfuerzo y rectitud. Gracias por tu paciencia, impulso y, sobre todo, por tu apoyo en el camino que he emprendido, sé que te parece extraño y a pesar de eso estás a mi lado. A mi padre, quien fue mi ejemplo en el amor a las letras. Siempre estás conmigo aún si en ocasiones sólo es en espíritu. A ti debo mucho, lo que me enseñaste se puede ver incluso en esta tesis.

A mi hermano. Eres un ejemplo de fuerza de voluntad y gracias a ti nunca me he sentido sola.

A mi asesor y maestro, el Dr. Gerardo Sierra Martínez. Por permitirme ser parte del Grupo de Ingeniería Lingüística, por enseñarme lo que es la lingüística de corpus, por tenerme paciencia durante el proceso de elaboración de esta tesis y por la libertad con que se trabaja a su lado.

A mi co-asesor, el Dr. Juan-Manuel Torres-Moreno. Por enseñarme que los números nos dicen mucho, sólo hay que saber escucharlos. Por el apoyo que me brindó al diseñar el programa que calcula la similitud semántica en esta tesis.

A la Dra. Iria da Cunha Fanego. Mi maestra, mi amiga, mi co-asesora extraoficial y guía de vida. Sin ti no existiría esta tesis. Gracias por llevarme de la mano en este viaje académico. Espero que algo de tu rigor se encuentre en mi trabajo de aquí en adelante.

A la Mtra. Margarita Palacios Sierra. Por regalarme su tiempo al revisar esta tesis y por la amabilidad con que me presentó sus observaciones.

Al Dr. Alfonso Medina Urrea. Por la minuciosa revisión de esta tesis y el tiempo que dedicó en ello.

Al Mtro. Javier Cuétara Priede. Por acceder a ser parte de mi sínodo y porque en sus clases comenzó mi interés por la lingüística.

A Oscar Escamilla que me quiere y a quien quiero tanto. Por estar conmigo tanto tiempo a pesar de mi poca tolerancia al estrés y la larga lista de defectos que tengo. *Je t'aime mon ours stellaire!*

A mis compañeros del GIL, que no mencionaré por nombre pero que saben quiénes son y la gran amistad que nos une.

A la UNAM, gran casa de estudio que nos da tanto y nos pide tan poco.

1.Introducción

Una de las faltas intelectuales de mayor crecimiento en la actualidad es el plagio. Básicamente, se puede tomar como plagio el hecho de tomar el trabajo de alguien más y presentarlo como propio, por lo que existe una amplia variedad de trabajos que pueden ser víctimas de plagio.

En el ámbito académico actual ha incrementado la cantidad de plagios textuales que se llevan a cabo. Esto sucede gracias al acceso rápido a la información que actualmente permite el internet y a las herramientas de procesamiento de textos.. Afortunadamente, las mismas herramientas que han facilitado a las personas llevar a cabo los plagios pueden ser utilizadas por individuos o instituciones para llevar a cabo la detección de los mismos.

Uno de los métodos utilizados para ayudar a la detección de plagio es el cálculo de la similitud textual. Este método es un ejemplo de tantos que se han beneficiado de las herramientas computacionales de hoy en día. Es dentro de este ámbito que se inserta la presente tesis. En este trabajo se presentará un método innovador para el cálculo de similitud textual.

1.1. Motivación del trabajo

Actualmente, cada vez más personas tienen acceso a internet. Este aumento presenta una ventaja para los usuarios, ya que cada vez es más fácil el acceso y la distribución de información. Es así que una persona puede colocar sus textos u otras obras de su creación al alcance del público mundial gracias a la ventaja que representa el internet.

Pero con el creciente uso de internet como fuente y modo de acceso a la información, y gracias también a la ayuda que proporcionan los programas de procesamiento de textos, se ha hecho más fácil llevar a cabo actos de plagio. Es así que "En las últimas dos décadas se ha observado un crecimiento importante en los casos de plagio, sobre todo el de tipo académico" (Barrón-Cedeño, *et al*., 2010).

Era de esperarse, entonces, que este aumento en los casos de plagio despertara la preocupación de la comunidad académica. Esta preocupación, a su vez, ha levantado el interés por desarrollar métodos que ayuden a la detección de plagio.

Puesto de manera simplificada, actualmente las aproximaciones a la detección de similitud textual pueden ponerse en tres categorías: comparación basada en palabras, búsqueda en línea de párrafos característicos mediante motores de búsqueda y análisis estilístico (Maurer, *et al*., 2006). Estos métodos no representan una ayuda suficiente puesto que todo texto posee estructuras localizadas en varios niveles y, para poder realizar un análisis más fino, es necesario emplear métodos que posibiliten el estudio de las estructuras a todos estos niveles.

Esta tesis se origina al observar la posibilidad de aportar a la detección de plagio mediante la creación de un método que contemple tanto la estructura discursiva de un texto como su contenido léxico-semántico.

1.2. Hipótesis

Esta investigación se realiza principalmente con base en las hipótesis siguientes:

- Las estructuras discursivas de un texto original y un texto que lo parafrasee han de ser similares, aunque el léxico y la estructura sintáctica sean diferentes.

- Ha de ser posible establecer la similitud entre el contenido semántico de un texto original y el de un texto que lo parafrasee mediante un cálculo.

Las hipótesis secundarias pertenecientes a esta tesis son las que siguen:

- Con la comparación de las estructuras discursivas de dos textos ha de ser posible identificar las unidades discursivas que son similares entre un texto y otro.
- A pesar de que las estructuras de dos textos pertenecientes al mismo género discursivo pudieran ser parecidas, la estructura discursiva de un texto que parafrasea a otro (tomado como original) presentará mayor parecido con éste que la estructura de otro texto que no representa una paráfrasis con el mismo original.
- Un texto que parafrasee a otro habrá de presentar mayor similitud semántica con éste que un texto que no lo haga.

De constatarse las hipótesis planteadas, el análisis de la estructura discursiva y el cálculo de similitud semántica entre textos podrían ser considerados como parámetros adicionales para la detección de similitud textual.

1.3. Objetivo

Esta tesis persigue los siguientes objetivos:

- Comparar las estructuras discursivas de un corpus de textos originales y de textos parafraseados a distintos niveles (bajo y alto), para observar si estas estructuras discursivas son similares.

- Calcular la similitud semántica[1] entre los textos originales y los textos parafraseados a distintos niveles pertenecientes a un corpus, para observar si el resultado obtenido es útil para la detección de similitud textual.
- Comprobar si la comparación de las estructuras discursivas de textos originales y textos que los parafraseen, junto con el cálculo de su similitud semántica pueden resultar de utilidad para la detección de similitud textual.

1.4. Estructura de la tesis

Esta tesis se divide en cinco apartados principales además de la presente introducción:

En el capítulo 2 se presenta un breve estado del arte respecto a los métodos automáticos actuales de detección de plagio. Primero contiene una introducción a la detección de plagio y luego se presentan los métodos existentes para llevarla a cabo.

En el capítulo 3 se describe la *Rhetorical Structure Theory* y su funcionamiento. Primero se presenta lo que para este trabajo se tiene como análisis del discurso y luego se describe su relación con esta teoría.

En el capítulo 4 se presenta el recurso de semántica léxica llamado EuroWordNet. Primero se describen las características del recurso llamado WordNet del que deriva EuroWordNet para después describir las especificidades de éste.

En el capítulo 5 se describe la metodología que se siguió y se discuten los resultados obtenidos. Se describe en este capítulo cada uno de los pasos de la metodología propuesta para luego discutir los resultados obtenidos.

Finalmente, en el capítulo 6 se presentan las conclusiones a las que se llegó respecto a los objetivos e hipótesis planteados.

[1] En el capítulo 5.4.1 (que tiene comienzo en la página 78) se explica el método utilizado para el cálculo de similitud semántica inspirado del algoritmo de Maynard (1999).

2.Detección de plagio

En este capítulo se presenta el estado del arte respecto a la detección de plagio. Para esto, primero se hará una introducción a lo que es la detección de plagio para después hablar de los métodos que se utilizan para llevarlo a cabo.

2.1. Introducción a la detección de plagio

El reúso de texto se define como la actividad mediante la cual el material escrito preexistente se reutiliza en la creación de un nuevo texto (Clough, 2003 p. 1). Esta actividad no es algo reciente. Es muy común que en la investigación y en la literatura se lleve a cabo el reúso de textos. En el primer caso, es necesario para sustentar las ideas nuevas y, en el segundo caso, suele tomarse como homenaje a otros autores. Pero el reúso de texto no siempre es aceptable.

Actualmente nos encontramos ante un problema en crecimiento, el plagio. Para este trabajo, dado que no tiene como finalidad encontrar una definición de lo que representa el plagio, se tomará como referencia la definición de Clough (2000) "when the work of someone else is reproduced without acknowledging the source"[2]. Según Barrón-Cedeño *et al*. (2010) "Distintos objetos pueden ser plagiados; desde un fragmento de texto o un programa informático hasta una fotografía, pintura o pieza musical". En este trabajo nos interesa el problema del plagio en textos.

Gracias al creciente uso del internet como fuente de información y gracias a las ventajas que otorgan los programas de procesamiento de textos cada vez es más fácil llevar

[2] Cuando el trabajo de alguien más es reproducido sin reconocer la fuente. (La traducción es de la autora)

a cabo actos de plagio. El aumento en la cantidad de plagios que se llevan a cabo actualmente ha llamado la atención de la comunidad académica y, dado que no se ha podido evitar este problema, se ha dado un gran interés en la creación de métodos para detectarlo.

2.2. Herramientas y métodos existentes

Barrón-Cedeño *et al*. (2010) mencionan que "Dada la enorme cantidad de documentos existentes, y disponibles, la detección manual de plagio resulta imposible. Por ello, es necesario el desarrollo de herramientas computacionales que asistan al ser humano en esta tarea: los llamados detectores automáticos de plagio".

Actualmente existe una gran variedad de herramientas para la detección de plagio. Para presentarlas, éstas se pueden dividir de varias maneras: atendiendo al tipo de plagio que se busca, según el método que utilizan, dependiendo de los idiomas en los que funcionan o teniendo en cuenta el número de textos que se analizan. En este trabajo se hablará de las herramientas para la detección de plagio dividiéndolas del último modo. Cabe mencionar aquí que las herramientas de las que se habla a continuación, así como la mayoría de herramientas para la detección de plagio, fueron creadas para su uso con textos en inglés.

Tomando en cuenta lo anterior, básicamente se puede hablar de dos maneras de detectar el plagio en textos. Ambos métodos requieren una cantidad distinta de documentos para analizar y ambas arrojan resultados de una naturaleza diferente. La primera manera es la detección intrínseca de plagio, la cual requiere solamente del texto en el cual se sospecha puede haber plagio. La segunda manera es la detección extrínseca de plagio, en la cual es

necesario tener al menos un documento con el cual comparar el documento sospechoso de plagio. A continuación se describe de manera más extensa ambos métodos.

2.3. Detección intrínseca

La gran cantidad de textos a los que se tiene acceso y que facilitan los plagios es lo mismo que dificulta la comparación de documentos si no se tiene idea de cuál podría ser el supuesto documento original del cual se reutiliza el texto. La detección intrínseca de plagio se creó pensando en esta situación, donde se tendría que comparar un documento con una cantidad enorme de textos para encontrar el origen u orígenes del supuesto plagio. Entonces, la detección intrínseca de plagio se hace con la intención de determinar si hay la posibilidad de plagio en un texto determinado sin buscar el origen del supuesto plagio.

Un ejemplo de detección intrínseca de plagio puede ser mediante la estilometría (Rosas González, 2011). Lo que se hace es detectar si hay una continuidad (coherencia) en el estilo a lo largo de un texto. A pesar de que el tema restringe hasta cierto punto el tratamiento de un texto, cada autor tiene una manera particular de expresarse. Las particularidades del estilo de un autor se pueden observar en el uso de las palabras y signos de puntuación, y la longitud de las oraciones. De esto se aprovecha la estilometría para detectar el plagio. En todo texto escrito por un solo autor, a lo largo del mismo se verán las marcas de estilo del que lo ha escrito. Si se encontrara que en alguna parte del documento esas marcas difieren, se podría suponer que esa parte del documento no fue escrita por el mismo autor.

Pero no siempre es necesario un análisis profundo para detectar fragmentos sospechosos de plagio en un texto. Por ejemplo, si en un texto de escritura informal se observara un cambio repentino a estilo formal, es posible sospechar que ese fragmento no

haya sido escrito por el mismo autor sin necesidad de hacer uso de conteos de longitud de oraciones u otros recursos estilométricos.

El principal problema de estos métodos es que, dado que no se compara el texto sospechoso de plagio con ningún otro texto, no es posible demostrar la existencia de un plagio como tal, ya que no se busca el posible documento plagiado. A pesar de esta desventaja, detectar el plagio de manera intrínseca posee la ventaja de que no es necesario tener ningún texto para comparar con el supuesto plagiario y, por lo mismo, sólo se analiza un texto acelerando el tiempo del análisis.

2.4. Detección extrínseca

La detección extrínseca de plagio se hace comparando el supuesto plagio con al menos un texto. En esta búsqueda, lo que se intenta detectar es la existencia de similitud entre textos. Básicamente se puede decir que, entre más se parezcan dos documentos o incluso fragmentos de documentos, más probabilidad hay de que uno sea plagio del otro.

Debe mencionarse ahora que, para que se lleve a cabo una verdadera detección de plagio, en todos los casos es necesario un experto que dictamine si existe o no plagio en un texto dado. Es por esto que en lo que resta de este trabajo se hará referencia a los métodos utilizados para la llamada detección extrínseca de plagio como métodos de detección de similitud textual.

Ahora bien, para hablar de los métodos de detección de similitud textual "conviene girar la vista hacia una clasificación que tome como base el tipo de operaciones realizadas al texto reusado" (Barrón-Cedeño, *et al*., 2010). A continuación se utilizará la tipología

propuesta por Maurer *et al*. (2006)[3] para hablar de los métodos de detección de similitud textual.

Copia exacta.

En este caso se reutiliza el texto tal cual se encuentra en el documento original. Es por esto que la forma más básica de detectar este reúso es mediante la búsqueda "manual" de frases características de un texto en internet por medio de los motores de búsqueda disponibles. En este caso, es necesario identificar previamente los fragmentos de texto que resulten sospechosos de ser tomados de otros textos.

Cuando se lleva a cabo un plagio mediante copia exacta del texto original es probable que haya incoherencias entre el texto general y el fragmento o fragmentos plagiados, lo que facilita la identificación de frases sospechosas mediante la lectura del supuesto plagio. Pero en los casos en los que el documento sospechoso es muy extenso o se tiene un grupo de ellos, es posible hacer uso de alguna herramienta de detección intrínseca de plagio para detectar los fragmentos que probablemente no pertenezcan al autor del texto sospechoso. Habiendo identificado los fragmentos sospechosos, se lleva a cabo una búsqueda "manual" de éstos en la red con lo cual se comprueba la existencia de este tipo de plagio al encontrar un documento fuente que contenga el mismo fragmento textual. En este caso, lo que se busca es una coincidencia exacta entre fragmentos de texto o incluso de un texto completo.

Sin embargo, para cuando se reúsa texto exactamente como se encuentra en el documento original, el método más efectivo para detectarlo es el llamado *fingerprinting.* En este método, el supuesto plagio se separa en cadenas textuales o "fingerprints" y éstas se comparan con las de otros documentos en búsqueda de similitudes entre textos. Este

[3] Como se describe en Barrón-Cedeño *et al*. (2010)

método de detección de similitud textual se utiliza tanto en herramientas en línea como en programas descargables de ejecución local.

Clough (2000) menciona Plagiarism.com como el servicio de detección de similitud textual en línea más grande disponible. Este servicio actualmente se deriva en dos herramientas para la detección y cálculo de similitud textual. La primera herramienta es llamada iThenticate[4]. Esta herramienta permite al usuario cargar un documento para que el sistema lo compare con documentos localizados en dos fuentes diferentes: páginas en internet y publicaciones y artículos académicos localizados en su base de datos. Este sistema permite también crear un repositorio de comparación personalizado según las necesidades del usuario. iThenticate arroja entonces varios reportes de "similaridad". En estos reportes se observa el porcentaje de similitud encontrado entre textos y la ubicación en internet de los textos con los que se encontró más similitud.

La segunda herramienta provista por Plagiarism.com es llamada OriginalityCheck, la cual se encuentra disponible por medio de Turnitin[5], un derivado de Plagiarism.com. Esta herramienta está dirigida principalmente al ámbito escolar. Este sistema compara un documento enviado por el usuario con documentos localizados en su base de datos. Estos documentos se obtienen de tres sitios diferentes: páginas web, documentos previamente enviados para su análisis con la herramienta y libros y publicaciones periódicas en la base de datos ProQuest[6]. Esta herramienta arroja un "reporte de originalidad" en el cual el usuario puede ver, sobre el texto analizado, en qué partes del texto se encontraron coincidencias con los documentos de la base de datos de la herramienta. En el mismo

[4] http://www.ithenticate.com/
[5] http://www.turnitinadmissions.com/
[6] http://www.proquest.com/

reporte de originalidad se muestra una lista de los documentos en los que se encontraron similitudes ordenados por porcentaje de similitud.

Otro ejemplo de herramienta en línea para detección de similitud textual es Docoloc[7]. Esta herramienta utiliza las capacidades de búsqueda y clasificación de la interfaz de programación de aplicaciones o API de Google. Esto significa que, en este caso, se busca similitud con documentos encontrados en internet. Este sistema proporciona al usuario la capacidad de elegir las fuentes de internet de las que se obtienen los documentos a comparar. La ventaja de esto es que se hace posible reducir el número de resultados obtenidos con bajo nivel de similitud. La desventaja en este caso es que al reducir la cantidad de documentos analizados, se puede llegar a reducir la cantidad de documentos fuente identificados.

Las herramientas en línea tienen como ventajas principales que no se necesita descarga ni instalación para utilizarlas y que pueden utilizarse en cualquier lugar del mundo. A pesar de esto, no necesariamente son herramientas gratuitas y varias de ellas exigen registro para poder utilizarlas. Por supuesto, toda herramienta en línea conlleva la desventaja de que debe tenerse acceso a internet para hacer uso de ellas.

La ventaja de utilizar los documentos que se encuentran en internet para buscar los posibles documentos originales de un documento supuestamente plagiado yace en que "In nearly all recent examples of copyright violations in scientific, academic and scholarly areas the original source of the plagiarized passages can be found on the Internet."[8] (Brandt, *et al*., 2010). Aún si internet no es la única fuente disponible para los plagiarios, supone una

[7] http://www.docoloc.de/

[8] En casi todos los ejemplos recientes de violaciones a derechos de autor en áreas científicas, académicas y escolares la fuente original de los pasajes plagiados se puede encontrar en internet. (La traducción es nuestra)

buena fuente de documentos de comparación para los sistemas de detección de plagio. Esto sucede en especial cuando no se tiene una idea previa del posible origen del supuesto plagio.

En el caso de los programas de ejecución local, un ejemplo muy común es WCopyfind[9]. Esta es una herramienta gratuita que, tras su descarga, compara desde un documento sospechoso hasta varios documentos a la vez. Para hacer la búsqueda de similitud textual este programa compara el documento o documentos sospechosos con un corpus de documentos localizados en la máquina donde se ejecuta. Esto implica que el usuario debe proporcionar los documentos de comparación para poder utilizar esta herramienta. Esta característica permite acelerar la búsqueda de concordancias entre "fingerprints", ya que se compara una cantidad menor de textos que en los sistemas mencionados anteriormente.

Sin embargo, este programa presenta ciertas desventajas porquerequiere que el usuario provea los posibles documentos fuente. Por una parte, el usuario debe tener una idea previa de los posibles orígenes del supuesto plagio además de la posesión de esos documentos en versión digital. Esto supone una búsqueda previa por parte del usuario del programa para localizar los posibles documentos fuente. Sin embargo, los documentos de los cuales se hace plagio no siempre se encuentran disponibles de manera gratuita en la red, por lo que, dado el caso, el usuario de este programa tendría que pagar para disponer de estos documentos. Sumado a esto, la posibilidad de encontrar el documento fuente de un plagio disminuye al reducir los documentos que se comparan con el supuesto plagio. Por lo que, de no poseer una versión digital del documento fuente de un plagio, no se podría detectar su similitud con el documento sospechoso mediante este sistema.

[9] http://plagiarism.phys.virginia.edu/

Lamentablemente, los métodos de *fingerprinting* sólo funcionan adecuadamente en casos de copia exacta ya que, debido al tipo de comparación que se realiza entre los documentos, estos sistemas dejan de detectar la similitud al hacer cualquier tipo de modificación en las "fingerprints".

Copia modificada

En este caso, antes de reutilizar el texto el plagiario le hace algunas modificaciones. Para detectar el plagio cuando se hace por medio de una copia modificada lo que se busca es detectar y calcular la similitud entre el texto sospechoso o una parte de él y el texto o textos que podrían ser la fuente del plagio.

Dado que en los plagios hechos de esta manera el texto plagiado no se reúsa tal cual se encuentra en el documento original, no es posible detectar la similitud textual por medio de métodos como los utilizados en el caso anterior. Para este caso, se deben utilizar métodos de detección de similitud textual que no sólo tomen en cuenta las palabras exactas del texto sospechoso, sino que también tomen en cuenta las posibles modificaciones que el supuesto plagiario pudo haber hecho al texto original entes de reutilizarlo.

Entre las variaciones que se le pueden hacer a un texto antes de reutilizarlo se cuentan desde cambiar una letra o una conjunción hasta el cambio de sintaxis y/o de léxico. Es por esto que el plagio mediante copia modificada es el que conlleva más dificultades al momento de detectar la similitud entre textos.

Existe una gran cantidad de métodos que se utilizan para la detección de similitud en textos que se sospecha de plagio con modificaciones. La ventaja de estos métodos es que también resultan útiles en el caso de la copia exacta, por lo que si se tiene un posible plagio es posible detectar similitudes entre textos, ya sean en coincidencias exactas o en

similitudes de tipos variados. Los métodos para medir la similitud en el caso de la copia con modificaciones utilizan una gran variedad de enfoques.

Un ejemplo de estos métodos son los que realizan comparación por *n*-gramas. Un *n*-grama es una secuencia de elementos en donde *n* representa una cantidad de elementos determinada. Estos elementos pueden ser de diversa naturaleza, pero para el caso de la detección de similitud textual los *n*-gramas representan secuencias de *n* palabras. La herramienta ya mencionada WCopyfind puede usarse para detectar similitud textual mediante este método. Para hacerlo, se le permite al usuario elegir el tamaño de *n*-gramas a comparar.

La desventaja de que el usuario elija el tamaño de los *n*-gramas es que debe ser éste quien piense en un *n*-grama lo suficientemente pequeño para no buscar copias exactas y lo suficientemente grande para no encontrar similitud entre textos que usan términos similares. El aumento en el tamaño del *n*-grama aumenta el riesgo de que ese conjunto de palabras no se encuentre ya que se buscan coincidencias exactas. La disminución en el tamaño del *n*-grama, por su parte, disminuye la posibilidad de encontrar un conjunto de *n*-gramas que representen plagio.

Otro método de detección de similitud textual para este tipo de plagio se describe en Kang *et al*. (2006). En ese trabajo se habla de la herramienta PPChecker que, según se describe, divide los textos en oraciones y calcula una "medida de superposición" entre textos. Para esta herramienta, la similitud textual se calcula a manera de "medida de superposición" ya que, según dicen sus creadores, no es suficiente saber qué tantos elementos comparten dos oraciones, sino que el orden en que se encuentran esos elementos determina un nivel más alto de similitud entre textos. Esta herramienta toma como base Wordnet (ver capítulo 5) en busca de sinónimos para ayudar al cálculo de superposiciones.

Dicen los creadores de esta herramienta que mediante la información que se obtiene a nivel oracional, es posible obtener una medida de similitud por párrafo, y posteriormente, por documento.

Otras herramientas útiles para detectar este tipo de plagio son SCAM (Shivakumar, *et al*., 1995) que detecta similitud textual basándose en la cantidad de palabras coincidentes entre textos, CopyCatch[10] que hace esto mismo pero toma en cuenta las palabras de aparición única en los textos y CHECK (Si, *et al*., 1997) que compara los textos utilizando palabras clave.

Plagio traducido

Otra manera de llevar a cabo un plagio es mediante el reúso de un texto o fragmentos de texto mediante su traducción. En este caso no es posible utilizar un método de detección de similitud textual como los anteriores, ya que se trata con variación entre idiomas. En este caso, uno de los métodos más sencillos para detectar similitud entre textos es mediante una traducción del texto sospechoso de plagio al idioma que se crea puede ser el original para posteriormente utilizar alguno de los métodos de detección de similitud textual. En este caso, dado que no existen las traducciones exactas, lo más aconsejable es utilizar los métodos que se utilizan para la detección de similitud para copias modificadas.

Según mencionan Barrón-Cedeño *et al*. (2010), debido tal vez a la complejidad que implica, en la actualidad apenas comienza a tomar interés la detección de similitud textual para este tipo de plagio.

[10] http://cflsoftware.com/?page_id=42

3.La RST en el análisis del discurso

3.1. Generalidades del análisis del discurso

Como se puede ver desde el título de la tesis y de este mismo capítulo, uno de los enfoques del análisis aquí presentado es mediante una perspectiva de discurso, o más específicamente, mediante análisis discursivo. En este capítulo, se sentarán las bases teóricas al respecto.

Para empezar, se ha de decir qué es discurso. Pues bien aquí enfrentamos el primer problema ya que, como menciona Maingueneau (1980 p. 15), "Contrariamente a lo que sucede en otros campos de la lingüística, el análisis del discurso tiene grandes dificultades para dominar su objeto." En otras palabras, este campo de la lingüística se encuentra con una gran dificultad al definir lo que es discurso. Entonces, atendiendo a esta problemática de encontrar una definición de discurso aplicable a la lingüística en general, me tomo la libertad de mencionar solamente la definición que más se acerca a lo que en esta tesis se va a entender como discurso. Pues bien, y a pesar de que la misma Maingueneau enlista al menos seis definiciones, aquí se tomará como referencia lo que dicen Calsamiglia y Tusón (2007 p. 1):

> Hablar de *discurso* es, ante todo, hablar de una práctica social, de una forma de acción entre las personas que se articula a partir del *uso lingüístico contextualizado*, ya sea oral o escrito (…) Nos referimos, pues, a cómo las formas lingüísticas se ponen en funcionamiento para construir formas de comunicación y de representación del mundo —real o imaginario —.

En otras palabras, el discurso es una práctica comunicativa que incluye no sólo el significado contenido en un texto oral o escrito, sino que abarca incluso los parámetros

cognitivos y socioculturales bajo los que se crea ese texto. Ahora bien, en palabras de Van Dijk:

> Tal como ocurre con la especialización en otras disciplinas, los analistas del discurso pueden concentrarse en un aspecto, nivel o dimensión del texto o la conversación o, incluso, en una clase general de discurso. (2000, pág. 27)

El mismo Van Dijk (2000, pág. 23) identifica "tres dimensiones principales" del discurso: "a) el *uso del lenguaje*; b) la *comunicación de creencias* (cognición) y c) la *interacción* en situaciones de índole social". Tomando en cuenta estas dimensiones, para Van Dijk (2000, pág. 23) "no es sorprendente que sean varias las disciplinas que participan de los estudios del discurso" tal es el caso de la lingüística que participa principalmente en el estudio de la dimensión del discurso marcada con el inciso "a". Ya que la presente tesis cae en los trabajos sobre el uso del lenguaje, a continuación se hará referencia a las propiedades del discurso dentro de esa dimensión.

Siguiendo a Van Dijk (2000, pág. 28), "el análisis del discurso puede comenzar por el análisis de un nivel de manifestaciones observables o *expresiones*, a saber, *sonidos* audibles y *marcas* visuales (cartas, figuras, colores, etc.)". Estas manifestaciones observables equivalen a la primera propiedad del discurso en su dimensión de uso del lenguaje.

Derivado de la distinción entre lengua oral y escrita encontramos, entonces, en el ámbito del discurso la división entre discurso oral y discurso escrito. Esta clasificación no sólo se basa en el hecho de que estos dos tipos de discurso se transmiten y se producen de manera distinta, sino que su respectivo estudio también se lleva a cabo de diferente manera al del otro.

En el discurso escrito, a diferencia del discurso oral, pueden los diversos estudiosos ubicar sus primeras apariciones en el tiempo, dado que posee una representación en la realidad que se puede percibir no sólo al momento de su realización sino, y normalmente es así, posterior a la producción de este tipo de discurso. Pero su persistencia en el tiempo es sólo una de las características que marcan la diferencia entre el discurso oral y el escrito. Por otra parte tenemos la diferencia de forma, en tanto a que ambos tipos de discurso tienen recursos visuales para apoyarse. El discurso oral tiene para esto las gesticulaciones del orador entre otros recursos visuales que se tengan a la mano en el momento y lugar en que se lleva a cabo el acto de habla; por su parte, el discurso escrito tiene como recurso visual cualquier cosa que pueda percibirse en el soporte en el que se encuentra escrito. Estas diferencias, entre otras, son claves en el momento de definir qué tipo de discurso queremos analizar y cómo lo vamos a hacer. En esta tesis sólo atenderemos a un tipo de discurso, al escrito.

Continuando con la descomposición que hace Van Dijk de los distintos niveles del discurso, se puede proseguir con la segunda propiedad del discurso u "orden y forma" en palabras del mismo Van Dijk. Dice éste al respecto que:

> Siguiendo los pasos de la gramática de la lengua, por ejemplo, cabe esperar que el análisis del discurso también preste atención a la *forma* abstracta de las oraciones que lo componen: el *orden* de las palabras, las frases o las cláusulas u otras propiedades que estudia la *sintaxis*. (Van Dijk, 2000 pág. 29)

Deteniéndonos aquí, podemos hacer mención de lo que Lope Blanch llamaba "análisis gramatical del discurso" (1987). Respecto a este análisis, Lope Blanch menciona que su intención es estudiar "las estructuras básicas del discurso en lengua española, atendiendo tanto a su modalidad literaria –ensayo y novela exclusivamente–, cuanto a su

realización oral –en sus niveles popular y culto.” (1987 p. 11). Según la metodología que plantea para hacer esto, descrita de manera general, se dividen los textos a analizar en unidades sintácticas para así observar sus estructuras enunciativas.

A pesar de las similitudes que posee el “análisis gramatical del discurso” de Lope Blanch respecto al análisis de la segunda propiedad discurso, debemos tomar en cuenta que dentro de la descripción que hace Van Dijk del análisis del discurso menciona que:

> A diferencia de los lingüistas tradicionales, sin embargo, los analistas del discurso van *más allá de la frontera de la oración* en este caso: estudian cómo influyen en la forma de las oraciones otras oraciones próximas en el texto o la conversación. (2000, pág. 30)

Es por esto que, si tomamos al análisis del discurso como lo describe Van Dijk, el análisis llevado a cabo por Lope Blanch no forma parte tal cual de los análisis del discurso. Es más, Van Dijk posiciona a la “gramática del discurso” dentro de las disciplinas que dieron paso al “surgimiento de los estudios del discurso”, no como una disciplina tal cual del análisis o, como él lo llama, estudios del discurso.

Ahora, reanudando la descripción de las propiedades del discurso, la tercera es lo que Van Dijk nombra “sentido”. Esta propiedad del discurso, según este autor, es analizada típicamente por la semántica. Van Dijk también menciona que, para caracterizar un término tan “escurridizo” y en favor de la caracterización de las propiedades del discurso, da a esta palabra un sentido abstracto acerca del cual menciona: “Los lingüistas suelen referirse a estos sentidos abstractos con la expresión *representaciones semánticas*.” (Van Dijk, 2000, pág. 31). Aclarando la relación entre la semántica y el análisis de la propiedad de “sentido” del discurso, Van Dijk dice:

> Mientras que la sintaxis del discurso se ocupa de la estructura formal de las oraciones, la semántica del discurso estudia, más bien, la estructura de las proposiciones[11], en especial las relaciones entre las proposiciones de un discurso. (2000, pág. 32)

A la cuarta propiedad del discurso Van Dijk nombra "estilo". Dada la dificultad que presenta dar una definición de lo que es "estilo", este autor la presenta en "términos de *varición*" (Van Dijk, 2000 pág. 34). Esto es, cuando se presentan diversas opciones de palabras para hacer referencia a una misma situación o persona, la elección de un determinado ítem léxico en un contexto determinado "se dice que estamos frente a características del estilo del discurso".

La cuarta propiedad o "dimensión" del discurso, según Van Dijk, es la retórica. Esta propiedad se puede definir brevemente como "las estructuras especiales del discurso que atraen la atención".

La quinta y última propiedad del discurso dentro de la dimensión del uso del lenguaje es la de las "estructuras formales globales, denominadas también *estructuras esquemáticas o superestructuras*." (Van Dijk, 2000, pág. 36). Según explica Renkema "Las superestructuras son esquemas convencionales que brindan el formato global para el contenido macroestructural del discurso" (1999, pág. 83).

Teniendo esto en cuenta, es posible definir qué es el análisis del discurso. Para esto, Calsamiglia y Tusón (2007 p. 13) dicen que:

> *El análisis del discurso* es un instrumento que permite entender las prácticas discursivas que se producen en todas las esferas de la vida social en las que el uso de la palabra –oral y escrita—forma parte de las actividades que en ellas se desarrollan.

Pero esto no nos dice explícitamente qué se puede entender por análisis del discurso. El análisis del discurso no es una sola teoría o técnica, sino que son las varias

[11] En este caso Van Dijk utiliza el término proposición para referirse a una cláusula u oración.

técnicas y perspectivas que se utilizan para estudiar las prácticas discursivas tomando en cuenta el contexto, la finalidad y la manera en que se produce un discurso. En fin, podríamos decir que el análisis del discurso tiene un enfoque principalmente descriptivo, pragmático y estructural.

Sabiendo que hay diferentes técnicas y teorías para el análisis del discurso, se puede, a partir de ahora, hacer referencia a la teoría en la que toma base esta tesis para llevar a cabo los objetivos que se plantea dentro del análisis discursivo, la RST.

3.2. Introducción a la RST

La *Rhetorical Structure Theory* (RST) es una teoría de análisis discursivo mediante la cual es posible caracterizar la estructura jerárquica de un texto. Fue creada en 1988 por William Mann y Sandra Thompson con fines computacionales. En esta caracterización se llevan a cabo la división de un texto en partes de importancia discursiva y la descripción de las relaciones y la jerarquía entre esas partes.

Atendiendo a la caracterización que se presenta en el apartado 3.1 podemos decir que la RST realiza sus análisis atendiendo a las propiedades de "sentido" y de "esquemas" del discurso. Es posible afirmar lo anterior si además tomamos en cuenta que ambas propiedades refieren a estructuras del discurso y la primera con particular enfoque a "las relaciones entre las proposiciones de un discurso". Pero el análisis con la RST no abarca todos los objetivos de lo que normalmente se describe como "análisis del discurso". Es por esto que, a lo largo de esta tesis, nos referiremos a la RST como una teoría de análisis discursivo y no de análisis del discurso como tal.

En un inicio, sus creadores diseñaron esta teoría con la finalidad de que, con la información obtenida de los análisis de textos de producción humana, se pudiera hacer

producción automática de textos. Pero gracias a los datos que arroja el análisis de textos con base en esta teoría y a su diseño enfocado a fines computacionales, ha tenido una variedad de usos para los lingüistas en el área del análisis del discurso.

Los mismos creadores de la teoría enumeran sus ventajas:

> It identifies hierarchic structure in text. It describes the relations between text parts in functional terms, identifying both the transition point of a relation and the extent of the items related. It provides comprehensive analyses rather than selective commentary. It is insensitive to text size, and has been applied to a wide variety of sizes of text. (Mann, *et al*., 1988 p. 243)

Debieron añadir además que esta teoría no es exclusiva de una lengua en particular, o sea que puede ser utilizada independientemente del idioma en el que se requiera, lo que implica una gran ventaja.

Hay que anotar aquí que el nombre de la RST se traduce como Teoría de la Estructura Retórica tomando el uso en inglés de la palabra "teoría", el cual, según el diccionario Collins (2006) del inglés, es "a system of rules, procedures, and assumptions used to produce a result"[12]. En este caso *theory* refiere a un modelo de análisis más que a una teoría como se entendería directamente en español. Además, aún si el nombre de esta teoría implica la retórica, su visión no se enfoca a ninguna definición específica, ya sea filosófica o lingüística, de lo que es la retórica, sino que refiere a lo que entendemos como discurso.

Ahora bien, todo aquel que quiera utilizar la RST debe saber que la aplicación de esta teoría presupone un texto coherente, ya que parte de los datos que arroja dependen de

[12] Un sistema de reglas, procedimientos y asunciones usadas para producir un resultado.

esta característica. Es decir, no se puede utilizar este tipo de análisis para un texto incoherente.

Para comprender mejor a lo que nos referimos con coherencia de un texto baste con ver la explicación que proponen en la página web oficial de la RST en español[13]. Para definir el término coherencia dicen ahí lo siguiente:

> (…) es la ausencia de secuencias ilógicas y de lagunas. En otras palabras, cada parte de un texto coherente tiene una función —una razón verosímil o aceptable para su existencia— evidente a los lectores, y, además, produce la impresión de que no le falta nada.

En palabras de Calsamiglia y Tusón, "Al tratarse de una **interacción diferida**, el texto debe contener las instrucciones necesarias para ser interpretado." (Calsamiglia Blancafort, *et al*., 2007 p. 63).

Es gracias a estas características que la RST parece ser de gran utilidad en investigaciones tales como la creación automática de textos, el resumen automático, la traducción automática o la evaluación de textos escritos por estudiantes.

Un analista (también llamado anotador[14]) realiza el análisis de los textos con base en esta teoría. Para hacerlo, entre otras cosas que se mencionan en el apartado 3.6, el analista debe dividir cada texto en sus unidades mínimas y anotar la relación jerárquica discursiva que mantienen esas unidades. Finalmente, hasta este momento no existe un método de anotación automática de textos con la RST, por lo que tanto en este trabajo como en otros basados en la RST debe realizarse de manera manual.

Esta situación se menciona en da Cunha *et al.* (2010) en donde se habla de los analizadores discursivos que se encuentran disponibles actualmente para el inglés, el

[13] http://www.sfu.ca/rst/08spanish/introduccion.html

[14] En este trabajo llamamos indistintamente analista o anotador al que lleva a cabo el trabajo de análisis discursivo, pero no necesariamente el anotador lleva a cabo el análisis discursivo completo.

japonés y el portugués de Brasil. En dicho trabajo (da Cunha, *et al*., 2010) se puede ver cómo se encuentra en desarrollo un segmentador discursivo para el español y se explica cómo la realización de éste es el primer paso para un analizador discursivo automático.

A continuación, en los puntos restantes de este apartado, se observarán los pasos a seguir y las consideraciones necesarias para la aplicación y comprensión de los resultados que arroja el análisis discursivo con la RST.

3.3. Unidades Discursivas Mínimas

Como ya habíamos mencionado antes, todo texto que se vaya a analizar mediante la RST debe segmentarse en sus Unidades Discursivas Mínimas o EDUs. Para esto, se debe tener conocimiento de las características de éstas. Debemos mencionar que las características de las EDUs pueden variar dependiendo de las delimitaciones que les quiera dar el analista. Las que nosotros utilizamos son las mismas que aparecen en da Cunha e Iruskieta (2010). A continuación las detallamos.

Para este análisis, la primera característica que deben cumplir las EDUs es que deben incluir un verbo, ya sea en forma conjugada, en infinitivo o en gerundio. Los participios no se consideran dado el tipo de significado que aportan. Tampoco se toman en cuenta para la segmentación los verbos adjetivados o sustantivados. La única excepción a esto se hace en el caso de los títulos, ya que rara vez tendrán verbo pero poseen importancia para el análisis con la RST. O como dijeran Calsamiglia y Tusón (2007 p. 86), los títulos son enunciados con fuerza retórica.

Una sola oración puede estar formada por varias EDUs, lo que nos indica que, a pesar de que las EDUs deben contener un verbo, la segmentación en EDUs no es exactamente igual que la segmentación oracional. En ocasiones se encuentran EDUs

incrustadas en otras, por lo que los fragmentos de la EDU en la que se inserta la otra se segmentan como si fueran dos EDUs y luego se unen con la relación Same-Unit. Como se muestra en el ejemplo siguiente[15] :

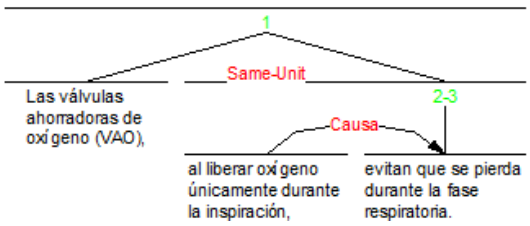


Figura 3.1 Ejemplo de unidad incrustada

Esta relación fue propuesta en Carlson y Marcu (2001) y tiene la función utilitaria de unir los fragmentos separados de una EDU en la que se inserta otra, por lo que, para la RST, no tiene significado discursivo y no se toma en cuenta para análisis posteriores.

Una oración de relativo, dado que completa el significado de otra oración, no se toma como una EDU individual sino que se toma como parte de la EDU cuyo significado completa. De la misma manera no se consideran EDUs las oraciones que forman el Objeto Directo o Indirecto de otra oración o que la adjetivan.

Las EDUs no siempre incluyen los llamados marcadores discursivos[16], por lo que estos no se toman como referencia para la segmentación. Finalmente, dado que los contenidos de paréntesis y marcas similares no siempre contienen un verbo, normalmente no se toman como EDUs a menos que cumplan con las características necesarias para serlo.

3.4. Estructuras jerárquicas

Ya dijimos que la RST nos proporciona información jerárquica del texto que analicemos con ella. En este punto, especificaremos el modo en que esto sucede.

[15] Ejemplo tomado de da Cunha e Iruskieta (2010).

[16] "Los *marcadores del discurso* son unidades lingüísticas invariables, no ejercen una función sintáctica en el marco de la predicación oracional y poseen un contenido coincidente en el discurso" (Portolés, 2007 p. 25)

En primer lugar, hay que decir que según la importancia que tenga una EDU dentro del discurso, y respecto a las EDUs con que se relacione, se caracterizará de dos modos:

Núcleo: Información relevante para los propósitos del autor.

Satélite: Información adicional que se proporciona sobre el núcleo.

Los grupos de EDUs relacionadas, también llamados SPANs, son susceptibles también de ser núcleo o satélite de otras EDUs o SPANs. Sabiendo esto, cabe mencionar que hay dos tipos básicos de estructuras jerárquicas que se observan en un texto, dependiendo de la relación que mantengan sus partes. La más sencilla de identificar es la estructura "nuclear". Esta estructura es la más común y siempre está formada por dos partes, un núcleo y un satélite. Hay ciertas relaciones que dan a las unidades discursivas esta estructura, pero más adelante se hablará de ellas. Como ya se mencionaba, de dos EDUs o SPANs que mantienen esta estructura, el núcleo mantiene la jerarquía más alta sobre la del satélite, y esto nos dice, en términos de intención del autor, que para el autor es más importante la información que se aporta en el núcleo que la que se aporta en el satélite.

La otra estructura es la multinuclear, como su nombre nos indica es la que contiene más de un núcleo. De esta manera es que se relacionan los elementos cuya importancia jerárquica es la misma. Puesto así, la estructura multinuclear por sí misma no contiene satélites, sólo núcleos. Esta estructura, a diferencia de la otra, no siempre está limitada a un número de elementos, las estructuras multinucleares generalmente pueden tener tantos elementos como el autor del texto los incluya.

3.5. Relaciones discursivas

Dado el tipo de jerarquía que denotan, hay dos tipos de relaciones, multinucleares y nucleares. Según la lista de relaciones en la que nos basaremos para este trabajo, tenemos

veintinueve relaciones discursivas en total: seis relaciones son de tipo multinuclear y veintitrés de tipo nuclear. Las describiremos a continuación:

3.5.1. Relaciones multinucleares

Las relaciones de este tipo en todos los casos estarán formadas por EDUs de igual importancia discursiva, por eso decimos que son de múltiples núcleos, dado que ninguna de sus partes es de menor importancia discursiva y, por lo tanto, sus componentes se consideran núcleos. Recordemos que la distinción "núcleo" o "satélite" se hace entre pares de EDUs atendiendo a la jerarquía de importancia que tiene una sobre la otra. Atendiendo a esta norma, en el caso en que ninguna de las EDUs posea menor peso discursivo que la otra, se dice que son núcleos. En los párrafos siguientes se explican las relaciones del tipo multinuclear. Estas explicaciones siguen el formato siguiente: nombre de la relación, caracterización de la relación y ejemplo de la misma[17].

Relación de Contraste[18]

Esta relación sólo se da entre pares de núcleos, o sea que no habrá relaciones multinucleares de Contraste que se den entre más de dos elementos discursivos (EDUs o SPANs). En este caso, ambos núcleos presentan una situación entre las que puede hacerse una comparación. Por esto, ambas situaciones se encuentran relacionadas dado que son comparables, pero contienen características que las hacen distintas y son objeto de comparación. Ejemplo:

[17] Todos los ejemplos de esta sección fueron tomados de da Cunha e Iruskieta (2010) y se muestran con la visualización que nos da la RSTtool.

[18] Las relaciones de la RST y su nomenclatura se propusieron originalmente por Mann y Thompson (1988). La traducción de la nomenclatura fue tomada de la página oficial de la RST (http://www.sfu.ca/rst/) y, tal como en inglés, los nombres de las relaciones comienzan con mayúscula.

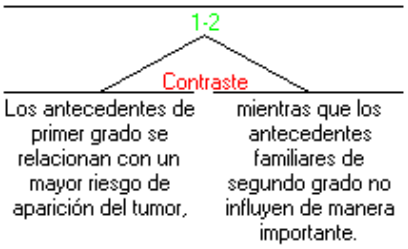


Figura 3.2 Ejemplo de relación de Contraste

Conjunción

En esta relación la intención del autor es de clasificar los núcleos como elementos de un mismo tipo que se unen y forman un conjunto. Comúnmente esta relación se da entre pares de núcleos pero no exclusivamente. Ejemplo:

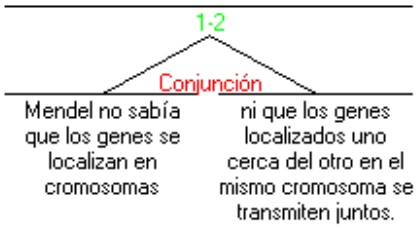


Figura 3.3 Ejemplo de relación de Conjunción

Disyunción

En esta relación los elementos que la conforman ofrecen dos alternativas para una misma situación. Los contenidos de los núcleos no siempre son mutuamente excluyentes, simplemente reflejan dos posibilidades para una misma cuestión. Ejemplo:

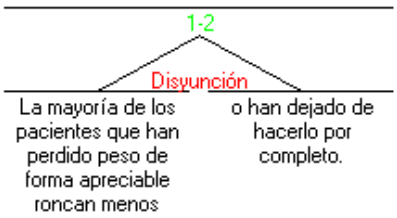


Figura 3.4 Ejemplo de relación de Disyunción

Lista

Los núcleos de esta relación son elementos de similares características que se encuentran redactados a manera de una lista. Esto no necesariamente implica que haya caracteres gráficos o marcadores discursivos que nos indiquen que es una lista de elementos. Los elementos unidos mediante la relación de lista no necesariamente se encuentran en un orden definido. La intención del autor aquí es simplemente la de enumerar ciertos elementos que considera son de un mismo tipo. Ejemplo:

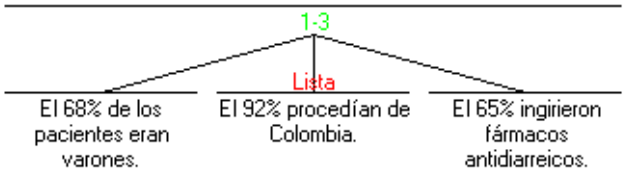


Figura 3.5 Ejemplo de relación de Lista

Secuencia

Para esta relación los elementos se encuentran de manera que forman una secuencia. Regularmente los núcleos de esta relación son situaciones que se realizan en cierto orden temporal y asimismo es como el autor las presenta. Ejemplo:

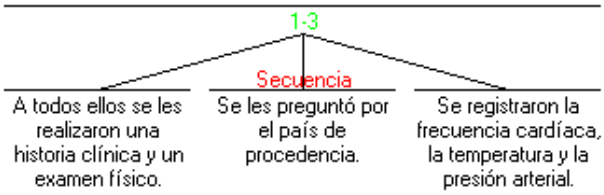


Figura 3.6 Ejemplo de relación de Secuencia

Unión

En este tipo de relación no hay restricciones en tanto a la cantidad o al tipo de núcleos, aquí la intención del autor es simplemente la de unir dos o más elementos. En ocasiones el autor relaciona ciertos elementos por medio de una Unión para así poder hacer referencia a ellos

en conjunto, pero esto no es una situación que se pueda generalizar. Otra característica de esta relación es que normalmente las EDUs que se relacionan de esta manera tienen el mismo sujeto. Ejemplo:

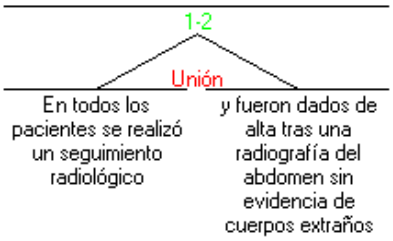


Figura 3.7 Ejemplo de relación de Unión

3.5.2. Relaciones nucleares

Las relaciones de este tipo, en todos los casos, estarán formadas por dos EDUs, de las cuales una será núcleo y la otra satélite. Este tipo de relaciones se dice que son nucleares dado que siempre habrá un solo núcleo del que dependerá el satélite. No debemos olvidar que este tipo de relaciones son binarias y que la denominación "núcleo" o "satélite" se le atribuye a una EDU con respecto a otra. Pensemos que en un texto todos los elementos discursivos se relacionan de alguna manera y por esto una unidad discursiva que sea núcleo respecto a otra puede ser considerada, a la vez, satélite con relación a una tercera EDU. En los párrafos siguientes se explican las relaciones del tipo nuclear. Estas explicaciones siguen el mismo formato que las de tipo multinuclear: nombre de la relación, caracterización de la relación y ejemplo de la misma.

Alternativa

En esta relación el núcleo es una situación condicionada por otra que se encuentra en el satélite. Esta condición supone que la realización de la situación del núcleo impide la realización de lo descrito en el satélite y viceversa.

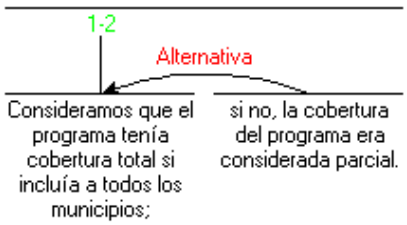


Figura 3.8 Ejemplo de relación de Alternativa

Antítesis

En este caso, el núcleo aporta cierta información o situación que se realiza o se supone verdad a pesar del dato o situación descrita por el satélite.

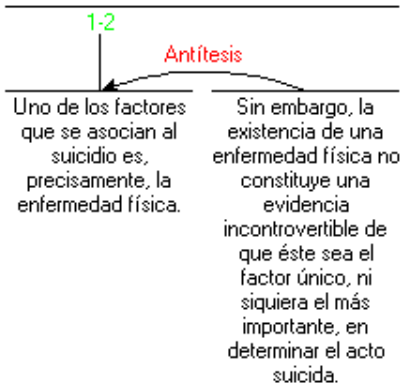


Figura 3.9 Ejemplo de relación de Antítesis

Capacitación

Para esta relación, el contenido del satélite da información que es necesaria para que se lleve a cabo la acción descrita en el núcleo.

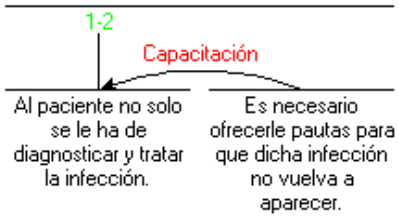


Figura 3.10 Ejemplo de relación de Capacitación

Causa

Cuando existe esta relación entre dos EDUs, se observa que el núcleo es una acción o situación que encuentra su origen en lo que describe el contenido del satélite. Es decir, la acción o situación mencionada en el satélite es el origen de la acción o situación descrita en el núcleo.

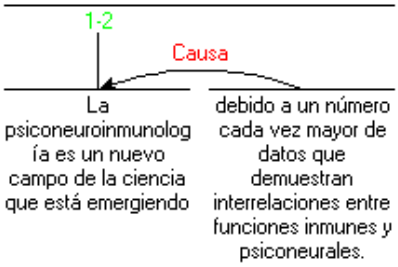


Figura 3.11 Ejemplo de relación de Causa

Circunstancia

El satélite en esta relación describe la situación contextual en la que se desarrolla lo descrito en el núcleo. Por tanto, el núcleo de esta relación contiene ciertos datos que encuentran su realización en tanto se desarrollan las circunstancias contenidas en el satélite.

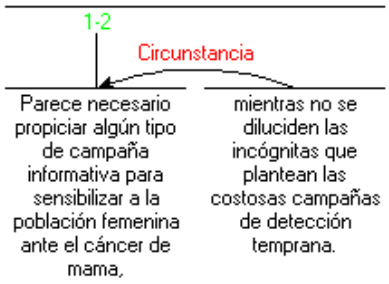


Figura 3.12 Ejemplo de relación de Circunstancia

Concesión

En este caso, el núcleo contiene una afirmación y se observa que la intención del autor es la de subrayar su afirmación mediante la inclusión de un satélite que aporta cierta información

que pareciera negar la validez de lo que se afirma en el núcleo, pero que realmente es complementaria.

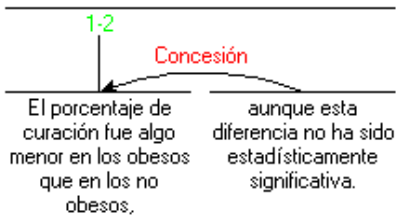


Figura 3.13 Ejemplo de relación de Concesión

Condición

El núcleo aquí contiene una situación hipotética y su satélite describe una situación de la cual depende la realización de lo descrito en el núcleo. Por tanto, si lo que se describe en el satélite no se llevara a cabo, lo descrito en el núcleo tampoco lo haría.

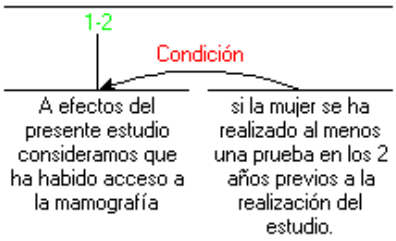


Figura 3.14 Ejemplo de relación de Condición

Elaboración

El satélite de esta relación presenta información adicional a lo descrito por el núcleo. Esta relación es la más común dado que, al hablar, la mayoría de las personas tienden a extender lo dicho mediante información adicional.

La mayoría de las relaciones de tipo nuclear se dan entre un núcleo que presenta cierta información y un satélite que presenta información adicional que se encuentra ahí por alguna razón en específico. En este caso, la razón de que las unidades discursivas que

representan un satélite de elaboración, respecto a un núcleo, se presenten en un texto es simplemente la de incluir información adicional.

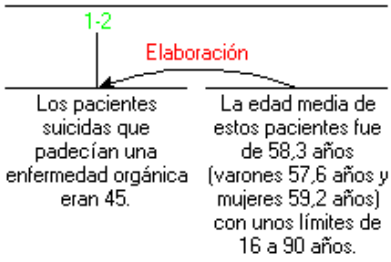


Figura 3.15 Ejemplo de relación de Elaboración

Evaluación

En este caso, el núcleo presenta cierta información y el satélite describe la opinión, ya sea positiva o negativa, del autor hacia la información contenida en el núcleo.

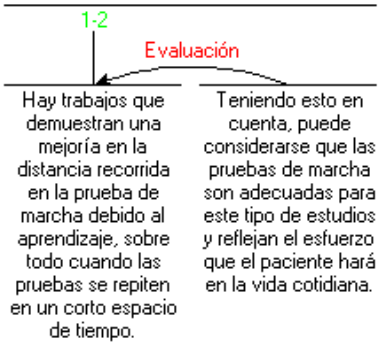


Figura 3.16 Ejemplo de relación de Evaluación

Evidencia

Esta relación se caracteriza por que su satélite presenta datos que aportan credibilidad a lo descrito en el núcleo.

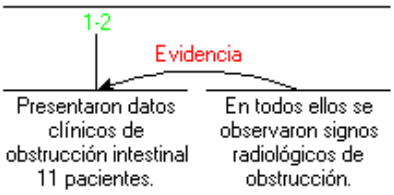


Figura 3.17 Ejemplo de relación de Evidencia

Fondo

En esta relación, el satélite aporta información sin la cual, el autor no cree que el lector pudiera comprender lo dicho en el núcleo. Por esto, en esta relación, el satélite siempre aparecerá antes que el núcleo.

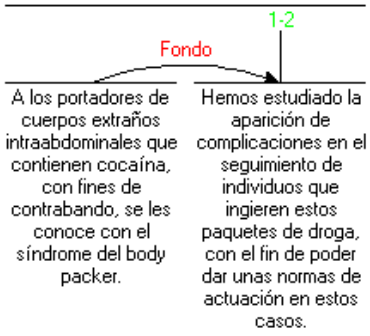


Figura 3.18 Ejemplo de relación de Fondo

Interpretación

En el satélite de esta relación se observan las conclusiones que el autor del texto presenta respecto a lo dicho en el núcleo. En esta relación, el autor no aporta una visión positiva o negativa respecto al núcleo.

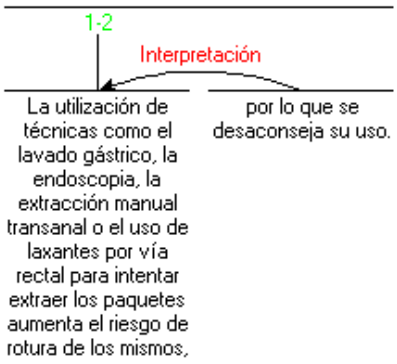


Figura 3.19 Ejemplo de relación de Interpretación

Justificación

Lo descrito en un satélite de esta relación representa la razón por la cual se presenta o se lleva a cabo lo descrito en su respectivo núcleo.

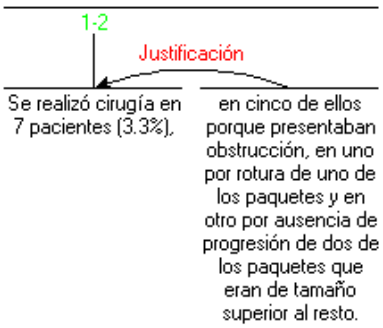


Figura 3.20 Ejemplo de relación de Justificación

Medio

En este caso el satélite presenta los métodos o herramientas que se utilizaron para la realización de lo descrito en el núcleo.

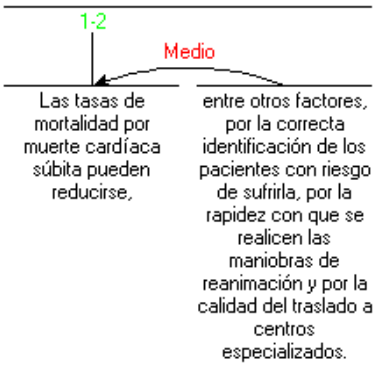


Figura 3.21 Ejemplo de relación de Medio

Motivación

En esta relación, el satélite presenta las razones por las cuales se llevó a cabo lo que se describe en el núcleo.

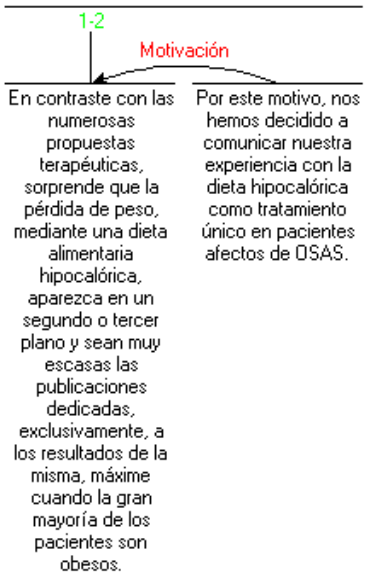


Figura 3.22 Ejemplo de relación de Motivación

Preparación

En esta relación el satélite es el título que presenta lo que a continuación se describirá en el núcleo. El satélite de preparación es la única EDU en la que se acepta que no exista un verbo, ya que el título mantiene cierta importancia discursiva dada su capacidad de resumir contenidos y atraer al lector.

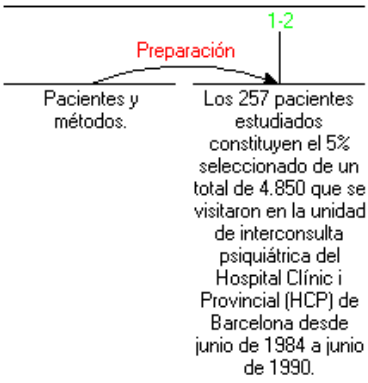


Figura 3.23 Ejemplo de relación de Preparación

Propósito

En el satélite de esta relación el autor presenta la finalidad que se tenía cuando se llevó a cabo la acción descrita en el núcleo. En el caso de que el núcleo contenga una situación o

acción que aún no se ha llevado a cabo, el satélite refleja las intenciones de aquel que lo llevará a cabo.

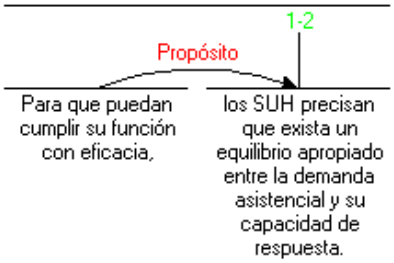


Figura 3.24 Ejemplo de relación de Propósito

Reformulación

En esta relación, el satélite, muchas veces de mayor extensión al núcleo, contiene la misma información que su correspondiente núcleo pero con otras palabras.

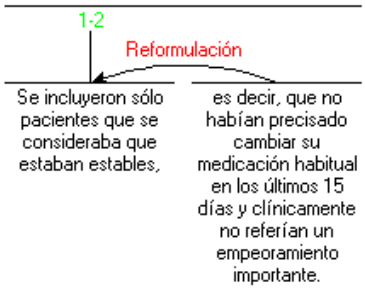


Figura 3.25 Ejemplo de relación de Reformulación

Resultado

Aquí el satélite presenta una situación proveniente de una acción o situación descrita en el núcleo. Es decir, dada la realización de lo descrito en el núcleo, sucede lo del satélite.

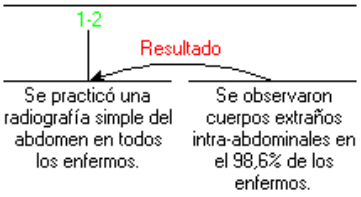


Figura 3.26 Ejemplo de relación de Resultado

Resumen

En este caso el núcleo no es una EDU sino un SPAN, o grupo de Unidades Discursivas Mínimas, y el satélite contiene la información presentada en el núcleo, pero de manera reducida.

Figura 3.27 Ejemplo de relación de Resumen

Solución

En este caso, el núcleo presenta un problema y en el satélite el autor presenta al lector una solución a ese problema.

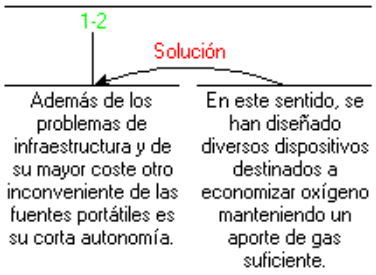


Figura 3.28 Ejemplo de relación de Solución

Unconditional[19]

En esta relación, el contenido del satélite describe una situación que el autor admite como aparentemente negativa de lo que se presenta en el núcleo, pero que se muestra como algo que en realidad no condiciona la realización de lo que se presenta en el núcleo.

[19] No-condicional.

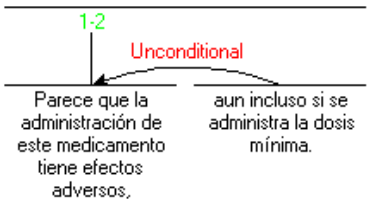


Figura 3.29 Ejemplo de relación de Unconditional

Unless[20]

En el núcleo de esta relación, el autor presenta una situación que se llevará a cabo solamente si el contenido del satélite no encontrara realización. El núcleo es una acción o situación que se encuentra condicionada de manera negativa por el satélite. En otras palabras, si sucede lo descrito en el satélite, no se llevará a cabo lo presentado en el núcleo.

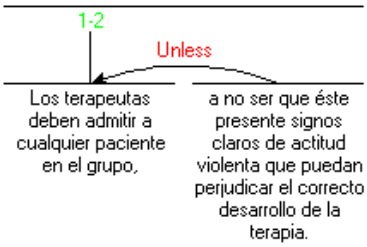


Figura 3.30 Ejemplo de relación de Unless

3.6. Participación del analista

El papel del analista dentro del análisis discursivo con la RST se puede dividir en seis actividades: segmentación discursiva, percepción del texto, detección de la intención del autor, detección de relaciones discursivas, construcción de árboles discursivos y análisis de los datos. Abundaremos en cada una a continuación.

[20] A menos que.

Segmentación discursiva

Para esta teoría y muchas otras de las utilizadas para el análisis discursivo, la primera parte del análisis es la identificación de las unidades de las que se compone el texto y la segmentación del mismo en esas unidades. En palabras de Casamiglia y Tusón (2007 p. 6):

> En definitiva, la complejidad que presenta cualquier pieza discursiva tiene que abordarse descubriendo en ella las unidades que constituyen sus diversas dimensiones (*módulos* para Roulet, *planos* para Adam y *niveles* para Viehweger, por ejemplo) que permiten su descripción y su posterior análisis de forma ordenada y sistemática.

Dentro de la RST las unidades que constituyen un texto son conocidas con el nombre de Unidades Discursivas Mínimas o EDUs por sus siglas en inglés. Las características de estas EDUs se mencionan en el apartado anterior. Es pertinente mencionar aquí que es trabajo del analista identificar todas las unidades que componen el texto que analiza y segmentar el texto en ellas. Para esto se puede hacer uso de una herramienta como la RSTtool, la cual se utilizó para hacer la segmentación de los textos que analizamos para esta tesis. Las EDUs no siempre son unidades discretas como podríamos pensar, por lo que el analista debe tomar en cuenta todos los elementos que debe contener una EDU antes de llevar a cabo su segmentación.

Percepción del texto por el analista.

En la creación de un texto, hay dos sujetos que se consideran los participantes de ese medio de comunicación, el autor y el receptor. El autor es el que plasma lo que quiere comunicar mediante texto, pero en todo caso considera un receptor, o sea alguien que leerá ese texto y para el cual debe ser comprensible. Tomando en cuenta el lector al que va dirigido un texto, su productor, idealmente, variará el contenido de modo que aquél lo reciba según sus

intenciones. Al llevarse a cabo un análisis con la RST tenemos una tercera influencia, el analista. Al momento de realizar la anotación del texto con base en esta teoría, y puesto que normalmente no se cuenta con la ayuda del autor del texto, el analista debe realizar ciertas suposiciones basadas en su propia percepción del texto. Aquí el analista tratará de averiguar la intención que pudo haber tenido el autor a la hora de redactar su texto aún si no se lo puede preguntar directamente.

Detección de la intención del autor.

Cuando cualquier persona escribe un texto, debemos presuponer que lo hace intencionalmente. Pero que no sólo es el acto de escritura lo que lleva una intencionalidad, sino que cada parte del discurso que se está creando tiene una razón de ser y una intención dentro del texto. Volvemos con esto a la necesidad de que un texto sea coherente para analizarlo mediante esta teoría. Es trabajo del analista, mientras hace su anotación, considerar cuál podría haber sido la intención del autor al colocar un fragmento de información con cierta forma y en cierto lugar dentro del discurso. La estructura que se busca obtener no es, en un principio, la estructura que el analista pueda observar en el discurso, sino la estructura que el autor le quiso dar al texto. Aplica de igual manera al momento de identificar el tipo de relación que hay entre las EDUs (ver 3.5) que forman el texto.

En este punto, el analista debe observar cómo se desarrolla el texto, pero así mismo debe tomar en cuenta el tipo de lector al que va dirigido el texto, ya que la intención del autor no sólo llega hasta el punto de tener coherencia en el texto, sino que, en ocasiones, ciertos fragmentos de información se incluyen en el texto con la intención de que su contenido y forma tengan influencia en el lector, ya sea convenciéndolo de algo u

ocasionando una reacción física en él. Este tipo de intenciones también son necesarias en el momento de identificar las relaciones entre las partes que componen el texto.

Detección de relaciones discursivas

Con el texto ya segmentado en sus unidades mínimas, se pasa a la detección de relaciones discursivas. En este punto, el anotador identifica de manera binaria la relación entre las EDUs, habiendo ya considerado los dos puntos previos y conociendo las características de las relaciones discursivas que se toman en cuenta en la RST. Este paso, si se hace uso de la RSTtool, normalmente se hace a la par que el paso que explicaremos a continuación.

Construcción de árboles discursivos

Éste, el paso final en la anotación de textos para la RST, es tal vez el más difícil, ya que para llevarlo a cabo el anotador debe tomar en cuenta lo dicho anteriormente y debe identificar la estructura general del texto y la estructura de las partes que la componen. Aquí es necesario mencionar que, habiendo identificado de manera binaria las relaciones discursivas y la manera en que se estructuran, también existe una relación y estructuración de los grupos de EDUs llamados SPANs, con lo que se observa la estructura del texto ya como conjunto de SPANs y en estos, a su vez, como conjuntos de EDUs.

Análisis de los datos

Ya con el o los árboles discursivos construidos, el analista puede pasar a la obtención de los datos que le interesan basándose en las características, que ahora son visibles en el o los textos anotados. Dependiendo de lo que el analista quiera obtener del análisis de sus textos, utilizará unos u otros datos de los que se muestran en la anotación.

3.7. La RSTtool como herramienta para el análisis con la RST

Actualmente, muchos de los usuarios de esta teoría, si no es que la mayoría, llevan a cabo sus análisis haciendo uso de una interfaz de anotación. Esto es porque, aún si las características de la RST nos permiten anotar y analizar los textos sin una interfaz para esto, se puede decir que el uso de una interfaz de anotación facilita la tarea del analista. Se facilita la tarea de anotación ya que una interfaz proporciona una imagen de los resultados de la anotación por medio de una estructura arbórea permitiendo así la visualización más clara de éstos. La herramienta más utilizada para la anotación de textos con la RST es la herramienta llamada RSTtool[21] (O'Donnell, 2000), misma que se utilizará en este trabajo.

La RSTtool posee la opción de abrir archivos previamente creados con esa misma interfaz los cuales tienen como extensión .rs2 o .rs3, dependiendo de la versión que tengamos de la herramienta. La que se utilizó en este trabajo fue la RSTtool 3.41, por lo que los archivos que obtuvimos tenían como extensión .rs3.

Cuando importamos un archivo de texto para su análisis con la RSTtool tenemos la capacidad de usar un master para la lista de relaciones que utilizaremos o de ir creando las etiquetas de las relaciones conforme trabajamos. La ventana en que se ejecuta la RSTtool tiene 4 pestañas en la que se observan interfaces diferentes, cada una con un fin específico:

[21] http://www.wagsoft.com/RSTTool/

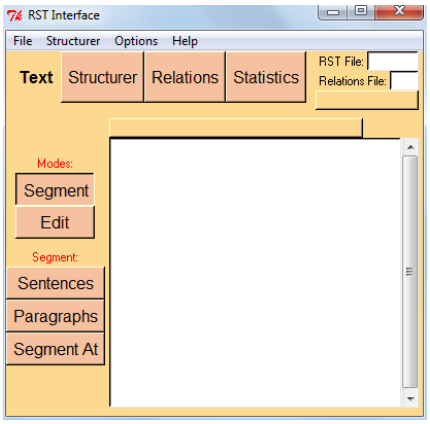


Figura 3.31 Vista de la RSTtool con la pestaña "Text" seleccionada

- Pestaña “Text”. Esta es la primera que aparece cuando importamos un archivo de texto para su etiquetado. En esta ventana aparece el texto que hemos importado y con hacer click esta interfaz añade una marca de segmentación. De no segmentar el texto, la herramienta considera el texto como un solo segmento. La herramienta nos permite también editar el texto directamente en la interfaz aunquehay que notar que cualquier cambio que aquí hagamos al texto sobre el que estamos trabajando no se verá reflejado en el archivo original de nuestro texto. Durante el proceso de estructuración también es posible regresar y hacer cambios en la segmentación o en el texto.
- Pestaña “Structurer”. En esta vista, la herramienta nos muestra ordenados horizontalmente los fragmentos de texto así como los hemos segmentado. Aquí podemos arrastrar los segmentos discursivos y unirlos según el tipo de relación que haya entre ellos, formando así un “árbol” discursivo.
- Pestaña “Relations”. Aquí es en donde podemos revisar la nomenclatura de las relaciones que estamos utilizando, así como agregar o quitar etiquetas.

- Pestaña “Statistics”. Esta pestaña nos permite revisar las estadísticas de nuestro texto etiquetado. Podemos ver aquí los números y porcentajes de aparición de las relaciones que utilicemos.

Una de las ventajas que ofrece esta herramienta es el hecho de que podemos encontrar disponible la RSTtool de manera gratuita para que cualquiera que quiera utilizarla la instale y la descargue en su computadora. Además, la página web en la cual la podemos encontrar ofrece manuales para la utilización de la herramienta. Otra ventaja es que la RSTtool se puede utilizar en varios sistemas operativos tales como Windows, Macintosh, y otros basados en LINUX y UNIX. También tenemos que los “árboles” discursivos que generemos mediante esta herramienta pueden guardarse como imágenes, lo cual representa una ventaja si queremos mostrar nuestros resultados a cualquier persona que no tenga disponible la herramienta o si quisiéramos insertar la imagen en algún archivo de Word o similares.

Como desventaja tenemos que el etiquetado XML que arroja esta herramienta fue diseñado especialmente para ésta, por lo que no puede ser utilizado por otras herramientas computacionales a menos que éstas hayan sido creadas pensando en la RSTtool. También tenemos que, puesto que la RSTtool sólo admite archivos de texto sin formato, cualquier texto que queramos analizar con esta herramienta debe pasarse a ese tipo de archivo, lo que implica un uso de tiempo para pre-procesar el o los textos que tengamos para análisis.

4.EuroWordNet como recurso de la semántica léxica

Como se verá en el capítulo 5 de esta tesis, uno de los aspectos importantes de la metodología del trabajo aquí presentado requiere de la base de datos léxica multilingüe llamada EuroWordnet. Puesto que ésta es una extensión de WordNet[22], el presente capítulo estará enfocado a la descripción de esta última con una mención posterior de sus diferencias con EuroWordnet. Esta exposición parte, además, de la plataforma teórica presentada a continuación.

4.1. Introducción a la semántica léxica

Moreno Quibén (2008) parte del estudio del significado para definir la semántica léxica. Afirma que, el estudio del significado se divide en dos sub-disciplinas: la semántica composicional y la semántica léxica. Estos dos bloques, anota el autor, se diferencian principalmente por sus objetos y métodos de estudio.

Por una parte, el objeto de estudio primordial de la semántica composicional es el significado de la oración. Esta sub-disciplina construye el significado de oraciones "a partir de unidades más pequeñas mediante patrones bien definidos" (Moreno Quibén, 2008, pág. 3). Es decir, para conocer el significado contenido en una oración, bajo esta sub-disciplina,

[22] http://wordnet.princeton.edu

se divide el objeto de estudio en unidades más pequeñas y se estudia el significado de cada una de ellas. Con base en los significados de las unidades que componen una oración, la semántica composicional determina el significado de la oración completa.

Por otra parte, la semántica léxica centra su estudio en la palabra. Esta sub-disciplina no divide su objeto de estudio para analizarlo, sino que lo aísla y estudia el significado a partir de la palabra misma, tomando ésta como unidad mínima. Dado que el uso de "palabra" como término llevaría a ambigüedades, la semántica léxica se refiere a su unidad mínima como "unidad léxica".

La diferencia metodológica más clara entre semántica composicional y semántica léxica se encuentra en que la primera analiza el significado de oraciones a partir de sus componentes y la segunda analiza aisladamente el significado de las unidades léxicas. Las tareas que lleva a cabo la semántica léxica para llevar a cabo el estudio del significado de las unidades léxicas, tal como se describe en Moreno Quibén (2008, pág. 3), son las siguientes:

a) Caracterizar el significado de las palabras.
b) Dar cuenta de las relaciones de significado entre las palabras.
    - *hombre-persona*
    - *caballo-córcel*
c) Caracterizar los distintos tipos de significado.
    - *suspender-catear* Diferencia social
    - *fresa-frutilla* Diferencia geográfica
d) Variación contextual del significado. El significado de las palabras varía en mayor o menor medida según el contexto en el que aparezca.
    - Juan empezó la botella (empezó a beber)

- Juan empezó el libro (empezó a leer, a escribir, . . .)

e) El cambio semántico. Las palabras amplían sus significados y adquieren contenidos nuevos para adaptarse a las nuevas realidades. Por ejemplo, el componente del significado que hace referencia a la forma de los objetos permite una exitosa extensión de significado basada en la forma.

- *ratón* como roedor
- *ratón* como dispositivo

4.2. Bases psicolingüísticas de WordNet

La psicolingüística, como su nombre nos puede dar una pista, tiene sus raíces en la psicología y la lingüística. A pesar de esto, esta nueva disciplina no es la suma total de sus predecesoras. Según Frías Conde (2002, pág. 10) en la psicolingüística participan varias especialidades aparte de la lingüística y la psicología. Según dice este autor, esta disciplina se apoya en la biología, en tanto que toma al lenguaje como una función de un sistema neurofisiológico, y en la computación, puesto que la lengua es un sistema que requiere tanto conocimientos como representaciones y algoritmos.

Miller *et al*. (1993, pág. 1) describen la psicolingüística como "an interdisciplinary field of research concerned with the cognitive bases of linguistic competence"[23]. En otras palabras, la psicolingüística se basa en varias disciplinas para investigar cómo se organiza en el cerebro la información necesaria para poder comunicarnos en un cierto idioma.

Como en cualquiera de las dos disciplinas de las que proviene, la psicolingüística no se centra en una sola teoría. A continuación solamente se presentan las ideas de la psicolingüística que los creadores de WordNet toman como base para su construcción.

[23] "Un campo de investigación interdisciplinario interesado en las bases cognitivas de la competencia lingüística" La traducción fue hecha por la autora de este trabajo.

Tengamos entonces que, para que una persona pueda hacer uso del lenguaje, debe poseer, previamente, conocimiento de los sonidos y significados de las unidades léxicas que conforman el lenguaje en el que se quiere comunicar. Pero la manera en la que está organizada esa información es de una gran complejidad. "Lexical memory must be so organized that the sounds and the contextually appropriate meanings of thousands of different words can be retrieved from memory at rapid rates."[24] (Miller, Fellbaum, Kegl, & Miller, pág. 183)

En un diccionario, por ejemplo, las palabras y sus conceptos se organizan por medio de la alfabetización de la representación ortográfica de las palabras. Pensemos ahora que la unión entre forma léxica (w*ord form*) y significado léxico (w*ord meaning*) fuera siempre binaria, es decir, que a cada representación fonológica o gráfica de una palabra le correspondiera un solo significado y a la inversa. Si este fuera el caso, probablemente una organización alfabética sería suficiente para representar, al menos de una manera básica, la organización mental que tenemos del vocabulario.

Pero este no es el caso, en la realidad, el sistema de referencias entre formas léxicas y significados léxicos es mucho más complejo. Según Miller *et al*. (1993, pág. 4) los psicolingüistas han llevado a cabo una variedad de experimentos que desarrollan la teoría de que el lexicón mental se organiza como un sistema de herencias de significado léxico. Es entonces que WordNet representa un lexicón mental hipotético con base en una jerarquía de significados.

[24] "La memoria léxica debe estar tan organizada que los sonidos y los significados contextualmente apropiados de cientos de diferentes palabras puedan ser recuperadas de la memoria a grandes velocidades" La traducción fue hecha por la autora de este trabajo.

Con esto en mente, a continuación se describe la construcción de la base de datos léxica WordNet.

4.3. Definición y descripción de WordNet

WordNet es la base de datos léxica para el inglés creada en la Universidad de Princeton en 1985. Esta base de datos está disponible de manera gratuita, ya sea en línea o para descarga en un ordenador. Desde su creación WordNet ha sido utilizada por cientos de investigadores en una gran variedad de aplicaciones. Ha sido tal la popularidad de WordNet que en la actualidad se ha hecho una extensión multilingüe de esta base de datos léxica creada sólo para el inglés. Esta extensión llamada EuroWordNet se utiliza como herramienta en este trabajo como se explicará en el capítulo 5 de esta tesis.

Tomando en cuenta las bases teóricas descritas en los dos apartados anteriores, ahora es posible definir con más claridad lo que es WordNet. Según Moreno Quibén (2008, pág. 63) WordNet es "Un sistema de semántica léxica en línea construido según las teorías psicolingüística [sic] actuales". En otras palabras, WordNet es una base de datos léxica sistematizada cuya construcción se basa en las pautas teóricas de la semántica léxica y la psicolingüística.

Por un lado, WordNet comparte objeto de estudio con la semántica léxica, o sea, ambos estudian la unidad léxica y su significado. Además de tomar como objeto de estudio el mismo que la semántica léxica, los creadores de WordNet atienden mediante esta base de datos a dos de las tareas de la semántica léxica:

- Caracterizar el significado de las palabras.
- Dar cuenta de las relaciones de significado entre las palabras.

WordNet puede atender estas tareas de manera simultánea al tomar en cuenta los estudios de la psicolingüística que indican que el significado de las palabras es representado en la mente por medio de relaciones jerárquicas entre los significados de las unidades léxicas. Esto significa que, al dar cuenta de las relaciones de significado entre las unidades léxicas que contiene, WordNet caracteriza el significado de las mismas.

Entonces se puede observar que así como la semántica componencial divide su objeto de estudio para llevar a cabo su análisis WordNet se aproxima al significado de las unidades léxicas por medio de las partes que componen ese significado. Tomando esto en cuenta, dicen los creadores de WordNet (Fellbaum, 1999, pág. XVI) que lo que su base de datos lleva a cabo puede considerarse una forma de semántica léxica componencial.

En los apartados siguientes se explica la construcción y el funcionamiento de WordNet.

4.3.1. Estructura de WordNet

WordNet representa los significados léxicos de dos maneras: mediante relaciones sinonímicas y mediante relaciones semánticas entre conceptos. La primera, y se podría decir que ésta es la base de la construcción de WordNet, se lleva a cabo en la construcción de los *synonym sets* o *synsets*.

Un *synset* es grupo de palabras que, en ciertos contextos, pueden representar un concepto dado. Se puede decir entonces que, en una primera instancia, los elementos que componen un *synset* pueden referir al significado unos de otros. Esto implica simetría en la relación semántica que mantienen los miembros de un *synset*, ya que, cuando un elemento *a* de un *synset* es sinónimo de un elemento *b*, el elemento *b*, a su vez, es sinónimo del elemento *a*.

Además hay que mencionar que los *synsets* están formados por palabras dentro de la misma categoría sintáctica. Es decir, WordNet está construido, elementalmente, por *synsets* de sustantivos, verbos, adjetivos y adverbios. Dado que hay palabras que según el contexto en el que se encuentren caen en una u otra categoría gramatical, en WordNet se colocan estas palabras en los *synsets* correspondientes dentro de cada categoría a la que pertenecen. Esto implica que una sola forma léxica puede aparecer en más de un *synset* para representar su significado para cada contexto.

La segunda manera en que WordNet representa los significados léxicos es por medio de relaciones semánticas entre los conceptos que denotan los *synsets*. Actualmente WordNet se construye a partir de una gran variedad de relaciones semánticas. A pesar de esto, existen más relaciones semánticas de las que contiene esta base de datos. En teoría, una palabra se puede definir a partir de las otras con las que se relaciona y según la relación que hay entre éstas. Pero WordNet no posee las suficientes relaciones semánticas para que esto suceda. Es por esta razón que, en la mayoría de los casos, se incluye una glosa explicativa al lado de los *synsets*.

Dado que las relaciones semánticas entre conceptos son diferentes dependiendo de la categoría sintáctica a la que pertenecen los *synsets* que los representan, en WordNet se divide la representación del lexicón mental según la categoría sintáctica de los *synsets* sobre los que se construye. Para esto, WordNet está formado por tres estructuras distintas, una para sustantivos, otra para verbos y una más para adjetivos y adverbios. A continuación se presenta la manera en que se construye cada estructura en WordNet.

Sustantivos

La estructura semántica que contiene WordNet en tanto a sustantivos toma como elemento constitutivo los ya mencionados *synsets*. Si pensáramos en WordNet como una pared

metafórica, los *synsets* serían los ladrillos. Pero una pared no se sostiene si sólo está formada por ladrillos apilados, es entonces que para unir los ladrillos en esta pared metafórica son necesarias las ya mencionadas relaciones semánticas.

Ahora bien, cuatro relaciones de significado de las varias que contiene WordNet se pueden tomar como básicas para comprender su estructura: la sinonimia, la hiponimia-hiperonimia, la meronimia-holonimia y, por último, la antonimia.

Respecto a la sinonimia, anteriormente se ha explicado su importancia fundamental en la construcción de WordNet. Esta importancia se puede ver en el hecho de que los *synsets* constituyen una parte fundamental de la estructura de WordNet como elementos constitutivos.

Por su parte, la relación semántica que, podemos decir, le sigue a la sinonimia es la de hiponimia- hiperonimia. Esta relación semántica es de las más frecuentes en WordNet dado que los sustantivos se organizan mediante una estructura jerárquica entre los *synsets* y esta relación es básicamente jerárquica. La relación de hiponimia- hiperonimia une *synsets* que refieren a conceptos más generales con los que refieren a conceptos más específicos. La hiperonimia e hiponimia son relaciones complementarias puesto que, si un concepto *a* es hiperónimo de un concepto *b*, a la vez significa que el concepto *b* es hipónimo del concepto *a*.

Tomemos como ejemplo[25] el *synset* {furniture, piece_of_furniture} del cual uno de sus hipónimos es {bed}[26]. Esta relación de hiponimia implica a la vez que el concepto "mueble" es el hiperónimo de "cama". En términos de significado, la relación hiponimia-

[25] Tomado de http://wordnet.princeton.edu/

[26] Para fines de este ejemplo nos referiremos a los conceptos representados por estos *synsets* como "mueble" y "cama" respectivamente.

hiperonimia entre estos dos conceptos, representados por *synsets*, implica "cama" como un tipo de "mueble".

Para la construcción de las estructuras jerárquicas en WordNet, los sustantivos se dividen en una serie de jerarquías que derivan de varios conceptos llamados "primitivos semánticos" o "*unique beginners"*. Estos primitivos semánticos funcionan como los conceptos genéricos de los cuales derivan el resto de los conceptos contenidos por las redes de significado de WordNet. Esta división de los sustantivos en varias jerarquías además ayuda a separar los conceptos reflejados por los *synsets* en campos semánticos relativamente distintos.

A pesar de que una opción hubiera sido suponer un concepto principal bajo el cual se jerarquizarían todos los sustantivos, este tipo de organización supondría un concepto abstracto que abarcara todos los sustantivos que se pudieran agregar. Recordemos que en esta jerarquía los conceptos están unidos por relaciones hiponimia- hiperonimia y eso supone que el concepto "superior" o hiperónimo es una generalización cuyas características son heredadas por sus hipónimos.

Retomando las que podrían ser las relaciones semánticas principales en WordNet, la tercera relación de significado, de las aquí mencionadas, es la de la meronimia-holonimia o relación parte-todo. En este caso, una relación de meronimia-holonimia entre dos *synsets* implica que uno forma parte de las características del otro. Tomemos como ejemplo los conceptos "brazo" y "hueso". Si tomamos que "hueso" es un merónimo de "brazo", esto a su vez, significa que el concepto "hueso" implica parte del concepto de "brazo". En este caso, podemos afirmar que hueso es una parte del brazo. En la siguiente figura se muestra un ejemplo visual de las relaciones semánticas aquí explicadas.

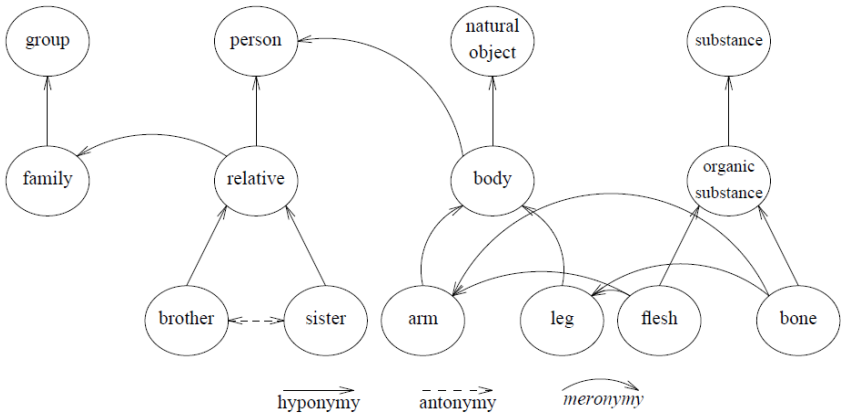


Figura 4.1 Representación de tres de las relaciones semánticas para los sustantivos en WordNet[27]

En esta figura es posible ver, entre otras cosas, cómo es que en inglés *brother* es hipónimo de *relative* y éste, a su vez, es merónimo de *family*. En otras palabras, el concepto de *brother* es un tipo de lo representado por el concepto *relative* y éste, por su parte, forma parte del concepto de *family.*

Por otra parte, y tomado ventaja de la Figura 4.1, es posible explicar además la relación de antonimia. De manera muy básica se puede definir esta relación como de oposición. Esta relación semántica, tal como la sinonimia, es simétrica en tanto que si las palabras *a* y *b* son antónimos, tanto *a* es antónimo de *b* como *b* lo es de *a*. Esta característica se puede ver en la Figura 4.1 Representación de tres de las relaciones semánticas para los sustantivos en WordNet, donde, así como *brother* es antónimo de *sister sister*, a su vez, es antónimo de *brother*.

Verbos

Al igual que los sustantivos, los verbos están ordenados mediante jerarquías y sus conceptos se representan mediante *synsets*. Pero la estructura jerárquica de los verbos, a diferencia de la de los sustantivos, se basa en varias relaciones semánticas de implicación. Dado que los significados verbales no pueden ser divididos en sus partes como los de los

[27] Tomado de Miller (1990 p. 25)

conceptos representados por sustantivos, los conceptos verbales se relacionan en WordNet mediante relaciones semánticas que denoten las implicaciones que contienen los conceptos verbales, ya sean relaciones en las que un concepto verbal represente una manera más o menos específica de llevar a cabo lo expresado por otro concepto verbal, relaciones en las que se implica una relación temporal entre dos conceptos verbales o relaciones en las que un concepto verbal implica la realización de lo denotado por otro concepto verbal.

La primera de estas relaciones semánticas es la denominada troponimia. En este caso, el tropónimo representa una caracterización más específica en el significado de los verbos. Tomemos como ejemplo aquí los conceptos verbales de "correr" y "transportarse". Para este caso, "correr" representa el tropónimo de "transportarse" puesto que correr es una manera específica en la que alguien se puede transportar. El significado que aporta un tropónimo depende del tipo de verbo que conecta. En el ejemplo anterior el tropónimo representa una manera específica de llevar a cabo el concepto con el que se relaciona.

Un ejemplo visual de cómo se llevan a cabo estas relaciones de implicación entre verbos se muestra en la figura siguiente.

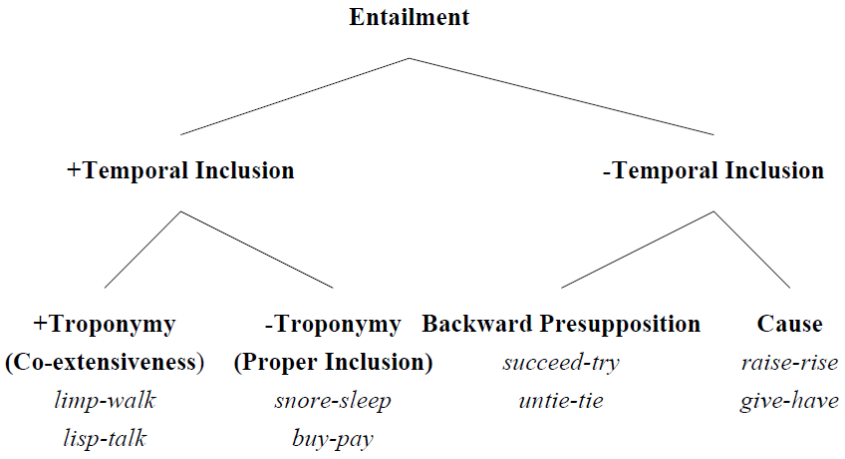


Figura 4.2 Ejemplo de relaciones de implicación entre verbos en WordNet[28]

[28] Tomado de Fellbaum (1990 p. 54)

En esta figura se muestra la relación entre varios pares de conceptos verbales. Se pueden ver en ella varios tipos de relaciones de implicación. Por ejemplo, la etiqueta +Troponimy para los conceptos verbales representados por *limp* y *walk* representa la implicación que el significado de *limp* hace del concepto representado por *walk*. Esto es, cuando se lleva a cabo lo representado por *limp*, por medio de esta relación se implica la realización de lo representado por *walk*. Puesto en español, "cojear" implica de alguna manera "caminar".

Adjetivos y adverbios

Finalmente, en WordNet los adjetivos y adverbios comparten una misma estructura dado que según se explica en la página web de WordNet, "There are only few adverbs in WordNet (hardly, mostly, really, etc.) as the majority of English adverbs are straightforwardly derived from adjectives via morphological affixation (surprisingly, strangely, etc.)"[29]

A diferencia de las estructuras que contiene WordNet para sustantivos y verbos, la base de la construcción de su estructura para los adjetivos y adverbios no es directamente por medio de la relación semántica de sinonimia, sino mediante una relación de antonimia. Para este caso, las relaciones semánticas entre conceptos no se muestran mediante jerarquías, sino mediante estructuras bipolares. Estas estructuras, como se muestra en la Figura 4.3, se forman por dos polos que se unen mediante una relación de antonimia entre sus respectivos núcleos. En cada polo el elemento principal o núcleo es un adjetivo o adverbio alrededor del cual se localizan otros elementos unidos a éste dada una similitud semántica entre los conceptos que representan.

[29] Solo hay unos pocos adverbios en WordNet (apenas, mayormente, realmente, etc.) puesto que la mayoría de los adverbios en inglés deriva directamente de adjetivos por medio de afijación morfológica (sorprendentemente, extrañamente, etc.)

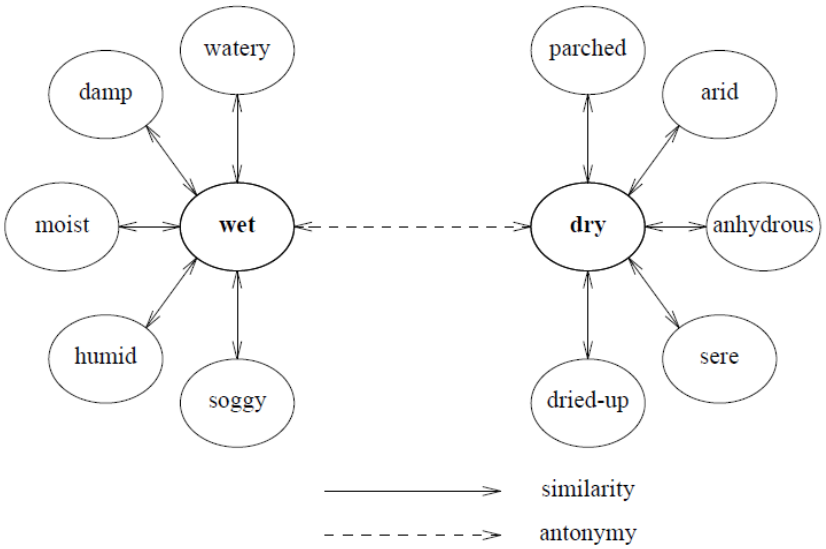


Figura 4.3 Ejemplo de estructura semántica entre adjetivos dentro de WordNet[30]

En esta figura es posible ver un ejemplo visual de la estructura en que se organizan los adjetivos y adverbios contenidos en WordNet. En este ejemplo, el núcleo es representado mediante los adjetivos *wet* y *dry* cuya relación semántica es mediante la antonimia. Además, esta figura nos muestra los adverbios con los que, en WordNet, se asemejan semánticamente estos dos adjetivos.

4.4. Diferencias entre EuroWordNet y WordNet

EuroWordNet, como ya habíamos mencionado, es una extensión multilingüe de la base de datos léxica WordNet. En el apartado anterior se muestra una caracterización generalizada de WordNet, pero a pesar de que EuroWordNet hereda las estructuras básicas de su antecedente contiene ciertas diferencias con éste.

EuroWordNet, a diferencia de WordNet, no es de carácter gratuito. Además, EuroWordNet no contiene datos semántico-léxicos de un solo idioma. EuroWordNet contiene información para al menos siete lenguas europeas. Y es dado este carácter

[30] Tomado de Fellbaum *et al.* (1990 p. 29)

multilingüe que se hicieron ciertos cambios estructurales a la base que proporciona WordNet.

A pesar de que EuroWordNet contiene datos de varias lenguas, los datos derivados de éstas se mantienen como una única base de datos en la cual las redes de significado se conectan para posibilitar la búsqueda inter-lengua de significados. Aún así, cada lengua posee su propia red de significado adaptada según lo necesitan las relaciones semánticas disponibles en ella.

Finalmente, hay que mencionar que para este trabajo se utilizó la versión en español de EuroWordNet, misma que se utiliza en Vivaldi *et al*. (2010). Según se menciona en Torres-Moreno (2011) esta versión de EuroWordNet contiene alrededor de 6,000 *synsets*, correspondientes a 10,000 variantes.

5.Detección de similitud mediante la RST y el cálculo de distancias semánticas

En general, nuestro método se puede dividir en dos partes atendiendo al enfoque teórico: la primera sería el análisis discursivo con la RST y la segunda, el cálculo de distancias semánticas. En específico, cuatro pasos forman nuestra metodología: Construcción de un corpus de paráfrasis, Anotación discursiva del corpus, Análisis manual y Análisis automático.

5.1. Construcción de un corpus de paráfrasis

Para comprobar nuestra hipótesis en este trabajo necesitábamos que nuestro corpus cumpliera con ciertas características:

- Que contuviera textos originales y sus paráfrasis.
- Que estuviera compuesto por textos cortos, dado que sería una sola persona la encargada de hacer la anotación y, puesto que aún no existe un anotador discursivo automático en español para la RST, de ser más extensos los textos, se intensifica la ardua tarea de hacer la anotación. Además, esta característica favorece la visualización de las estructuras discursivas anotadas.
- Que los textos fueran en español.
- Era preferible, además, que los textos abarcaran distintas áreas temáticas para tener cierta variedad en el corpus y que nuestros resultados no se ciñeran a un tema en particular.

Por esto, y dado el acceso libre que teníamos a él, preferimos partir de un corpus ya existente cuya creación se debe al Grupo de Ingeniería Lingüística. Aunque este corpus en su totalidad no cumplía con nuestros requisitos, decidimos tomar una sección, la cual cumplía con las características que necesitábamos y cuya temática atendía al origen del sushi, como base para crear nuestro propio corpus. El tamaño de ese fragmento, aún si sus características nos permitían utilizarlo para nuestros propósitos, no bastaba como corpus para obtener resultados suficientes.

Como se menciona en el artículo de da Cunha, Torres-Moreno y Sierra (2011), la construcción de un corpus anotado con la RST no es fácil. En todo caso, podría pensarse que es más difícil que la construcción de un corpus textual común, puesto que hay varios factores adicionales que se deben tomar en cuenta.

Para la construcción de nuestro corpus, se reunieron textos de diversas fuentes (Wikipedia, revistas científicas y periódicos) y temáticas (sushi, sexualidad y astronomía). Estos textos tendrían como tamaño de veinte a veinticinco "oraciones", tomando en este caso como "oración" el texto encontrado entre puntos, siendo indistinto si se trataba de punto y seguido o punto y aparte.

Posteriormente se solicitó a varios voluntarios (estudiantes de licenciatura, licenciados o doctores) que intencionalmente reformularan o parafrasearan dichos textos. Es necesario mencionar que para esta tesis se tomará como paráfrasis la reformulación de un texto conservando su contenido semántico pero variando el léxico, la sintaxis, la organización textual o discursiva o la cantidad de oraciones contenidas en el mismo. En este caso la reformulación se hizo en dos niveles:

Nivel bajo: Variación únicamente léxica.

Nivel alto: Variación léxica, sintáctica, de organización textual o discursiva y fusión o separación de oraciones.

Asimismo se buscaron textos que trataran sobre las mismas temáticas del corpus para comparar con los textos originales, y constatar que la posible coincidencia entre los textos originales y sus paráfrasis no sea fruto de una casualidad. Si hacemos una división atendiendo a las temáticas, nuestro corpus se divide en tres sub-corpus cuya temática respectiva sería sushi, sexualidad o astronomía. Cada uno de esos sub-corpus, a su vez, está formado por tres tipos de texto:

1. Textos originales

2. Textos parafraseados:

a) Nivel bajo

b) Nivel alto

3. Textos de la misma temática que los textos originales

Tenemos entonces como resultado tres sub-corpus formados por cuatro textos respectivamente, un original (OR), una paráfrasis de nivel alto (Pa), una paráfrasis de nivel bajo (Pb) y un texto de contraste, cuya temática, origen y extensión son similares a las del texto original (Pno).

Teniendo ya los textos que formarían el corpus, y antes de continuar con el siguiente paso de la metodología, se guardaron los textos en formato de archivo .txt y se eliminaron los títulos, ya que no se analizaría la similitud en ellos y no se le pidió a los participantes que hicieran una paráfrasis de los mismos. En tanto a extensión respecto a la cantidad de palabras, y habiendo ya eliminado el título de cada texto, puesto que no se tomaría en cuenta para el análisis, teníamos un corpus como sigue:

Tabla 5.1 Tamaño del corpus de paráfrasis

Corpus total		
	Textos	Palabras
Sushi	4	2513
Sexualidad	4	2436
Astronomía	4	3618
Total	12	8567

5.2. Anotación discursiva del corpus

Ya con un corpus que sirviera a nuestros propósitos, pasamos a la primera fase del método propuesto. Para comenzar a hacer un análisis basado en la RST es necesario tener el corpus anotado discursivamente, por lo que en esta fase se realizó la anotación discursiva de los tres sub-corpus a partir de la RST. Esta anotación se llevó a cabo siguiendo las pautas explicadas previamente en el capítulo 3. Cabe decir, entonces, que para realizarla se emplearon la interfaz RSTtool y la lista de relaciones discursivas ejemplificadas en el apartado 3.5. Pasada esta anotación discursiva, se tuvieron entonces doce árboles discursivos (como el que se muestra en la Figura 5.1) con un conteo de EDUs y de relaciones discursivas como se muestra en la Tabla 5.2.

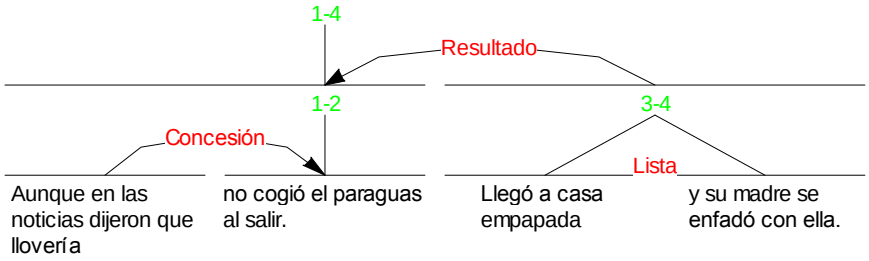


Figura 5.1 Ejemplo de árbol discursivo según la RST. Tomado de da Cunha *et al.* (2010 p. 2)

Tabla 5.2 Número de EDUs y relaciones que forman el corpus

Corpus total		
	EDUs	Relaciones
Sushi	183	165
Sexualidad	139	130
Astronomía	161	133
Total	483	428

En esta tabla podemos observar que no hay, como tal, una relación de proporción entre la cantidad de EDUs y de relaciones discursivas en los textos. Esto sucede dado que no todas las relaciones se dan entre pares de EDUs y además las unidades discursivas se pueden relacionar con más de una EDU a la vez. Éste último hecho se puede observar en la Figura 5.1, donde la segunda EDU es núcleo, a la vez, de una relación de Concesión y una de Resultado.

5.3. Análisis manual

El siguiente paso en nuestra metodología consiste en un análisis manual cuya realización describiremos a continuación.

Ya con el corpus anotado con la RST, comparamos las anotaciones para observar si había similitud entre ellas y, por consiguiente, similitud entre las estructuras discursivas de los textos. Se realizaron tres tipos de comparaciones entre los textos de la misma temática:

A) OR con Pa

B) OR con Pb

C) OR con Pno

Si tomamos en cuenta que nuestro corpus se compone por textos de tres temáticas, se hizo un total de nueve comparaciones. Esto sucede dado que se llevaron a cabo los tres tipos de comparaciones respectivamente para el sub-corpus de sushi, el sub-corpus de sexualidad y el sub-corpus de astronomía.

5.3.1. Discriminación de relaciones discursivas

Previo a los análisis específico y general dentro de esta primera aproximación a la detección de similitud textual con el análisis discursivo con la RST como una de sus bases, consideramos necesario definir qué relaciones discursivas analizaríamos para cada par de

textos en cada una de las nueve comparaciones. Para esto, se hicieron nueve tablas de discriminación de relaciones tales como la Tabla 5.3 que se muestra más adelante.

Para la construcción de estas tablas de discriminación de relaciones utilizamos como base las estadísticas que nos aporta la RSTtool de los textos previamente etiquetados. En esas estadísticas se contabiliza cada aparición de las relaciones de la RST dado un texto etiquetado. Es necesario señalar que para este trabajo elegimos que se contabilizara una vez por cada grupo de núcleos de una relación multinuclear. Esto se debe a que la RSTtool nos permite elegir de qué forma se cuenta cada relación multinuclear, ya sea con el método utilizado en este trabajo o una vez por cada EDU que se posea ese tipo de relación.

No todas las relaciones que aparecieran en los textos serían objeto de análisis. En primer lugar no se analizaron las relaciones multinucleares. Esto dado que el número de EDUs que forman este tipo de relaciones es variable. Y, ya que las comparaciones se harían entre parejas de textos, era necesario tener un número estable de EDUs por relación y este tipo de relaciones no lo tiene. Tampoco se tomaron en cuenta las relaciones de Elaboración. La razón para esto es que esta relación es la de uso más frecuente y por lo tanto tampoco es característica de algún discurso en particular[31]. Dado que ésta es una primera aproximación a la detección de similitud textual se creyó pertinente comparar sólo las relaciones que pudieran caracterizar un discurso para así evitar que la similitud encontrada fuera resultado de la casualidad. Por último, puesto que los análisis se harían a manera de comparaciones entre pares de textos y no es posible comparar relaciones que no se encuentren en ambos textos, no serían analizadas las relaciones que no aparecieran en el par de textos a comparar.

[31] Se puede ver la alta frecuencia de aparición de las relaciones de elaboración en los corpus existentes anotados con la RST.

Tomando esto en cuenta, las tablas de discriminación de relaciones se realizaron con el fin de destacar las relaciones que serían observadas al hacer las comparaciones. Una de las condiciones para que una relación discursiva fuera sujeto de análisis era su aparición en el par de textos a comparar. Por esta razón, era necesario hacer una discriminación que atendiera a las estadísticas por pares de textos. En adición, estas tablas servirían para tener una idea de cuántas de las relaciones totales que aparecen en los textos podrían ser objeto de comparación.

Se hizo una tabla de discriminación de relaciones, como la Tabla 5.3, para cada comparación. Para hacer esta tabla primero se extrajeron las estadísticas arrojadas por la RSTtool de los textos a comparar y se colocaron en paralelo. Después, las etiquetas y cantidades de las relaciones multinucleares presentes se movieron a la parte superior de la tabla. En la celda a la izquierda de éstas se colocó la etiqueta "relaciones multinucleares" para distinguirlas. A continuación, se colocaron los datos de las relaciones de elaboración debajo de las relaciones multinucleares. Se marcaron las relaciones de Elaboración mediante la etiqueta "relación de Elaboración", también en la celda a la izquierda de éstas.

Posterior a esto, se reunieron las relaciones que no aparecían en ambos textos y se colocaron debajo de las relaciones de Elaboración. Estas relaciones se marcaron mediante la etiqueta "relaciones únicas" dado que se encontraban sólo en uno de los dos textos a comparar. Con los datos ordenados de esta manera, quedaron juntas en la base de la tabla las relaciones que serían sujeto de análisis y se les etiquetó como "relaciones coincidentes".

Estas tablas de discriminación de relaciones incluyen la cantidad de apariciones de cada relación en su correspondiente texto. Gracias a este dato es posible observar el número probable de Unidades Discursivas Mínimas a analizar. Esta cantidad probable se hace

aparente si se toma en cuenta que las relaciones marcadas como “relaciones coincidentes” son siempre binarias.

Entonces, al multiplicar por dos la cantidad de apariciones de cada una de estas relaciones y después de sumar los resultados, se obtiene un estimado de la cantidad de EDUs que serían objeto de análisis en cada texto. Esta cantidad es una aproximación dado que una sola unidad discursiva puede relacionarse con más de una EDU.

Tabla 5.3 Comparación de estadísticas de dos textos del sub-corpus de sexualidad

	OR Sexualidad		Pa sexualidad	
	Relación	Cantidad	Relación	Cantidad
Relaciones multinucleares	Contraste	1	Contraste	1
	Disyunción	1		
	Lista	1		
	Secuencia	1	Secuencia	1
	Unión	3	Unión	2
Relación de Elaboración	Elaboración	16	Elaboración	13
Relaciones únicas	Causa	1		
			Circunstancia	1
			Conjunción	1
			Evaluación	1
			Medio	1
Relaciones coincidentes	Concesión	3	Concesión	1
	Fondo	1	Fondo	1
	Interpretación	3	Interpretación	4
	Resultado	2	Resultado	1

En esta tabla podemos observar que, de treinta y tres relaciones que aparecen en el texto original u OR del sub-corpus de sexualidad, nueve cumplen con los requisitos de los tipos de relaciones discursivas que compararíamos. Esto equivale al 27% del total de relaciones que se encuentran en ese texto. Asimismo, se observa que de veintiocho relaciones discursivas que hay en el texto de paráfrasis alta o Pa del mismo sub-corpus de sexualidad, siete serían sujeto de análisis. En otras palabras, se compararía el veinticinco por ciento de las relaciones de este texto.

Si observamos la Figura 5.2, podemos ver cuántas relaciones discursivas "comparables" y "no comparables", en promedio, aparecen en nuestro corpus etiquetado. Esto es, respectivamente, el promedio de relaciones que analizaríamos y el promedio de relaciones que no serían analizadas. En esta gráfica se observa la proporción de los promedios de relaciones "comparables" y "no comparables" por tipo de comparación.

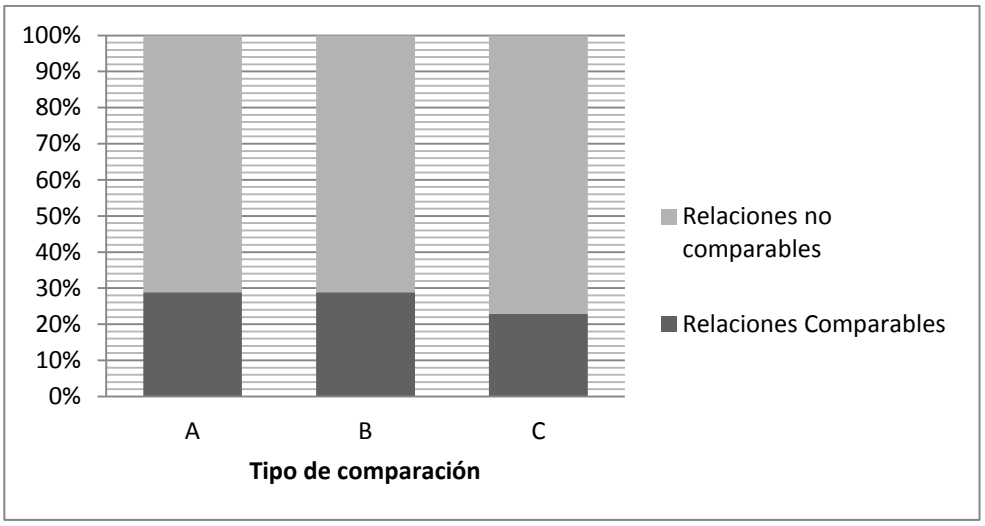


Figura 5.2 Relaciones comparables y no comparables resultado de las tablas de discriminación de relaciones

Pese a que esta discriminación aparentemente limita el alcance de nuestras comparaciones, hay que notar que el elevado porcentaje de relaciones que no se tomarán en cuenta se debe en gran parte a la alta frecuencia de uso de la relación de Elaboración, misma razón por la que no se tomó en cuenta para las comparaciones. Lo anterior se puede observar con más claridad en la Figura 5.3. Esta figura muestra la composición de las relaciones "no comparables" por cada tipo de comparación.

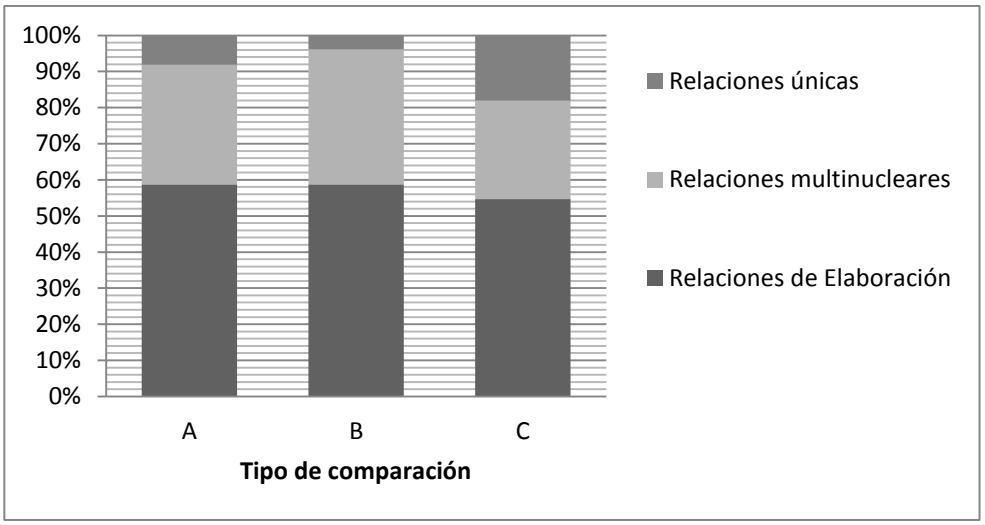


Figura 5.3 Composición de las relaciones no comparables

5.3.2. Análisis específico

Como ya se había mencionado, el segundo paso de nuestro método es la comparación de las anotaciones hechas con la RST a los textos de nuestro corpus. Esta comparación, por su parte, se divide en un análisis específico y un análisis general. Ahora explicaremos cómo se llevó a cabo el análisis específico.

En el análisis específico se hizo una comparación preliminar del contenido semántico de las EDUs cuyas relaciones discursivas resultaron sujeto de análisis en las tablas de discriminación de relaciones descritas en el apartado 5.3.1. Para hacer esto, lo primero que se hizo fue extraer de los textos etiquetados las EDUs cuyas relaciones discursivas fueran marcadas como "relaciones coincidentes" en las tablas de discriminación, por ejemplo la Tabla 5.3.

Para esta extracción no se tomó en cuenta las relaciones entre grupos de EDUs, también llamados SPANs. Esto se hizo dado que buscábamos que, como una primera aproximación, nuestro análisis atendiera a los contenidos y datos aportados por las EDUs individuales. De esta manera se puede ver qué datos aportan los elementos más básicos de

los árboles discursivos, aún si es posible que en trabajos posteriores se pudieran incluir elementos más complejos.

Se tomaron las EDUs extraídas y se ordenaron en tablas, como la Tabla 5.4, a las que nos referiremos como tablas de comparación específica. Recordemos que en esta etapa se llevaron a cabo tres tipos de comparaciones en los tres sub-corpus que forman nuestro corpus, por lo que se hicieron nueve tablas.

Para formar estas tablas de comparación específica primero se crearon tres columnas. Después, se colocaron los núcleos y satélites del texto original u OR en la columna del extremo izquierdo, los núcleos sobre su respectivo satélite. Cabe recordar que a partir de este paso solamente se trabajó con los núcleos y satélites cuya relación fuera una de las "relaciones coincidentes" en las tablas de discriminación de relaciones. Posterior a esto, en la columna central se anotó para cada EDU en la columna de la derecha si se trataba de un núcleo o un satélite. Esto se hizo no sólo por claridad sino también para mantener la jerarquía entre las EDUs. A continuación, en la columna del extremo izquierdo se marcó la relación discursiva mantenida entre los núcleos y satélites de la columna del extremo derecho. Con estos datos colocados, se procedió a agrupar los núcleos y satélites según su relación discursiva.

Hecho lo anterior, se crearon otras tres columnas a la derecha de las anteriores y se procedió de manera similar que con las tres columnas previas, sólo que en este caso se trabajó con los datos del texto con el que se compararía el original. El resultado fue una tabla de seis columnas, de las cuales las tres de la izquierda contienen los datos de texto original y las tres de la derecha contienen los datos del texto con el que se compararía el original. Para esclarecer esto, se procedió a colocar, en la parte superior de las respectivas

columnas, la etiqueta del tipo de texto del que se trataba la información, o sea OR, Pa, Pb o Pno.

Construidas estas tablas, en las mismas se comparó el contenido semántico de cada una de las EDUs de un texto con las del otro, tomando en cuenta que tuvieran la misma relación discursiva. Esta comparación se hizo, para descubrir aquellas unidades en las que el analista intuía coincidencia en el contenido. Posterior a esto, se ordenaron los datos de manera que quedaran a la par los núcleos y satélites que coincidieran en relación discursiva y en contenido textual. En caso de que ciertas EDUs no tuvieran "coincidente" a su lado se colocarían casillas vacías. Finalmente, se creó una nueva columna en el extremo izquierdo de cada tabla y se colocó un cero en la celda a la izquierda de las unidades con las que no se encontró coincidencia y un uno a la izquierda de las unidades en las que sí se encontró coincidencia. Con esto, se hizo posible observar y cuantificar más claramente el tipo y número de relaciones entre las EDUs con y sin coincidencia.

Dado que esta comparación se hizo como paso preliminar, se tomó en cuenta que posteriormente se haría una nueva comparación semántica por lo que en este caso 1 no representa un 100% de coincidencia semántica sino que se pone a manera de notación binaria en la que 1 significa "se encontró cieta coincidencia" y 0 significa "no se encontró coincidencia".

Tabla 5.4 Tabla de comparación específica. En estas tablas se lleva a cabo la comparación de EDUs

	Comparación B astronomía					
	OR			Pb		
	Relación discursiva		EDUs	Relación discursiva		EDUs
1	Propósito	Núcleo	El festival astronómico lleva como título “¡Haz Química con el Universo!”	Propósito	Núcleo	El evento lleva como título “¡Haz Química con el Universo!"
		Satélite	para unirse a las celebraciones del Año Internacional de la Química, declarado por la Asamblea General de la Organización de las Naciones Unidas (ONU) el 30 de diciembre de 2008, a propuesta de la Unión Internacional de Química Pura y Aplicada.		Satélite	para unirse a las celebraciones del Año Internacional de la Química, declarado por la Asamblea General de la ONU el 30 de diciembre de 2008, a propuesta de la Unión Internacional de Química Pura y Aplicada.
1	Resultado	Núcleo	La primera noche de las estrellas se realizó el 31 de enero de 2009 en 26 sitios arqueológicos y plazas públicas de 22 estados de la República.	Resultado	Núcleo	La primera vez se realizó el 31 de enero de 2009 en 26 sitios arqueológicos y plazas públicas de 22 estados de la República Mexicana.
		Satélite	Esta celebración logró convocar a más de 210,000 personas (el doble de la capacidad del Estadio Azteca);		Satélite	Este evento logró convocar a más de 210,000 personas (que corresponde al doble de la capacidad del Estadio Azteca);
0	Resultado	Núcleo	La segunda noche de las estrellas se realizó el 17 de abril de 2010 con el tema “Nuestro Universo en Movimiento”	Resultado	Núcleo	
		Satélite	reuniendo en esa ocasión más de mil 320 telescopios en 31 sedes		Satélite	

5.3.3. Análisis general

Como ya habíamos mencionado, posterior a la realización del análisis específico se hizo un análisis general, ahora explicaremos cómo se llevó a cabo.

Gracias al análisis específico que se hizo es posible identificar, en cada comparación, las EDUs que coinciden en tanto a relación y contenido. En el análisis general, se contabilizaron los resultados del análisis específico para así poder observar un panorama de los resultados obtenidos.

Para llevar a cabo este análisis, nos basamos en los valores que se colocaron en la columna de la extrema izquierda de las tablas de comparación específica (Ver 5.3.2). Se contabilizaron las coincidencias y diferencias que se encontraron para cada comparación de cada una de las temáticas y se colocaron estos datos a manera de tablas. Para hacer este análisis general, se construyeron dos tipos de tablas. Las tablas del primer tipo contienen el conteo de coincidencias y de diferencias para cada relación discursiva y las tablas del

segundo tipo muestran la cantidad de coincidencias y diferencias por cada comparación que se realizó.

A las tablas del primer tipo, como la Tabla 5.5, nos referiremos como tablas de coincidencias por relación. Estas tablas están formadas por cuatro columnas. Para construir estas tablas primero se colocó en la columna de la extrema izquierda la nomenclatura del tipo de comparación del que se muestran los datos, o sea A, B o C. Después, en la columna siguiente se especificó la relación discursiva de la cual se muestra la cantidad de coincidencias y diferencias encontradas. Posterior a esto, en la tercera columna hacia la derecha se colocó la cantidad de coincidencias obtenidas en las tablas de comparación específica para cada una de las relaciones discursivas. Y, finalmente, en la columna del extremo derecho colocamos la cantidad de diferencias que se encontraron para estas mismas relaciones. Para nuestro corpus se hicieron nueve tablas de coincidencias por relación, una para cada comparación de cada sub-corpus.

Las tablas de coincidencias por relación nos permiten ver en qué relaciones podemos encontrar el mayor número de paráfrasis identificables en tanto a coincidencia de relación y comparación del contenido. Gracias a estas tablas se puede observar qué relaciones aportan más coincidencias dentro de nuestro corpus.

Tabla 5.5 Coincidencias por relación

Sub-corpus Sexualidad			
A	Relación	Coincidencias	Diferencias
	Concesión	0	2
	Interpretación	1	1
	Resultado	0	2

Como ejemplo, la Tabla 5.5 nos muestra que la mayor cantidad de coincidencias dentro de la comparación de tipo A en el sub-corpus de sexualidad se encontraron para la

relación de Interpretación y, por su parte, la mayor cantidad de diferencias se encontró tanto en la relación de Concesión como en la relación de Resultado.

Al segundo tipo de tablas que se hicieron, como la Tabla 5.6, nos referiremos como tablas de coincidencias por comparación. Estas tablas están formadas por cuatro columnas también. En la columna de la extrema izquierda se encuentra, como en las tablas de coincidencias por relación, la nomenclatura del tipo de comparación del que se muestran los datos. En la columna siguiente, por claridad, colocamos la notación de los textos que se compararon. En la tercera columna hacia la derecha se colocó el total de coincidencias obtenidas en las tablas de coincidencias por relación para cada comparación. Finalmente, en la última columna se dispuso el total de diferencias encontradas por cada comparación. En este caso se hicieron tres tablas como la Tabla 5.6, una para cada sub-corpus analizado.

Tabla 5.6 Coincidencias por comparación

Sub-corpus Sexualidad			
		Coincidencias	Diferencias
A	OR – Pa	1	5
B	OR – Pb	5	5
C	OR – Pno	0	7

Mediante estas tablas de coincidencias por comparación podemos ver en qué tipo de comparación se encontraron más coincidencias y diferencias en tanto a relación discursiva y contenido. La Tabla 5.6, por ejemplo, nos muestra que por una parte, para el sub-corpus de sexualidad, la comparación tipo B es la que arroja más coincidencias y por otra parte la mayor cantidad de diferencias se encontró en la comparación tipo C.

5.4. Análisis automático

Para continuar con nuestra metodología se llevó a cabo un análisis automático cuya realización describiremos a continuación.

Para comprobar los resultados que obtuvimos del análisis manual y medir una similitud entre los fragmentos de texto identificados como coincidentes en el mismo análisis manual hicimos un cálculo de similitud semántica. Para realizar este cálculo se utilizó un programa en perl hecho por Juan-Manuel Torres-Moreno. El algoritmo que sigue este programa parte del trabajo de Vivaldi *et al.* (2010). Además, se utilizaron varios recursos del sistema de resumen automático CORTEX (Torres-Moreno, *et al*., 2001) para el pre-procesamiento del texto. Específicamente se utilizaron las listas de palabras funcionales y de reagrupamiento de familias léxicas de este sistema.

5.4.1. Algoritmo para el cálculo de similitud semántica

La idea general del algoritmo es calcular un valor de similitud semántica para los núcleos y satélites que se marcaron como coincidentes en las tablas de comparación de unidades discursivas mínimas descritas en el punto 5.3.2.

La Figura 5.4 muestra un esquema de la arquitectura general del algoritmo. Descrito de manera general, primero, las EDUs a comparar, tanto núcleos como satélites, pasan por un pre-procesamiento. Después se calcula la similitud semántica entre las palabras que forman las EDUs pre-procesadas. Las EDUS son representadas en forma de bolsa de palabras. La similitud entre pares de palabras se acumula en un valor único. Finalmente, se obtiene un valor final dividiendo el valor acumulado entre el número de parejas posibles. Este valor final se encuentra normalizado entre 0 y 1.

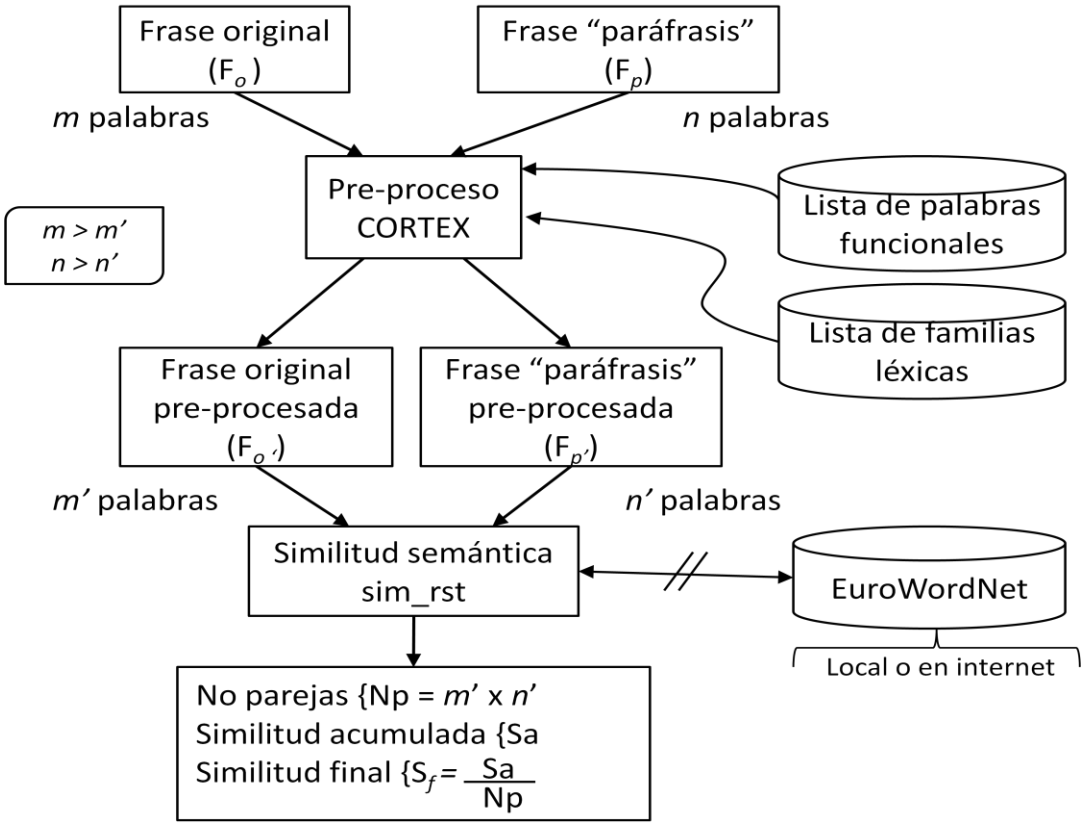


Figura 5.4 Arquitectura general del algoritmo de similitud semántica.

El proceso comienza con los dos fragmentos de texto o EDUs a compararse, o sea una EDU del texto original u F_o y una EDU del texto con el que se compara el original o F_p. Estas EDUs son los núcleos y satélites marcados como coincidentes en las tablas de comparación (ver 5.3.2). La frase original contiene *m* palabras y *n* son las palabras de la frase con la que se compara o "paráfrasis".

El primer paso de este algoritmo es un pre-procesamiento de F_o y de F_p. Primero se pasan ambas EDUs por un filtrado para remover signos de puntuación. Después se normalizan F_o y F_p a minúsculas. Se eliminan las palabras funcionales de ambas EDUs usando las listas de CORTEX y, finalmente, se reagrupan las palabras en familias léxicas, también usando los recursos de CORTEX. Con esto, tenemos como resultado los fragmentos de texto $F_{o'}$ y $F_{p'}$.

Este reagrupamiento de palabras por familia léxica es una alternativa más simple a la lematización. Como mencionan Barrón, Vila y Rosso (2010) "en un marco más realista, f debe compararse contra millones de f' s para generar una lista ordenada con base en las similitudes estimadas." Hacer millones de comparaciones de palabras no resulta práctico ni eficiente. A pesar de que la lematización disminuye la cantidad de palabras que se comparan, esa disminución no resulta tan significativa como con la agrupación de las palabras en familias léxicas.

Tomemos como ejemplo las siguientes palabras: imitadora, imitadores, imitando, imita, imitara. Si se hiciera una lematización de éstas se llegaría a los lemas "imitador" e "imitar". En cambio, si se agruparan por su familia léxica, tendríamos como único resultado "imitar". Para el grupo de palabras de este ejemplo se obtuvieron dos lemas pero solo un representante de la familia léxica. Puesto así, no parece haber una gran diferencia pero debemos tener en cuenta que esto significa que con la lematización se obtuvo el doble de palabras que con la reagrupación por familias léxicas. Tomemos también en cuenta que si se estuviera en una situación real de discurso probablemente serían más lemas, a diferencia de la agrupación por familias léxicas que daría el mismo resultado.

La lematización básicamente "es la transformación de las palabras en una expresión cercana a su raíz, llamada lema" (Sánchez Vega, 2011). Comúnmente se lleva a cabo en el pre-proceso de los sistemas utilizados en el procesamiento del lenguaje natural para analizar las palabras sin que los resultados resulten afectados por la derivación de palabras. Puesto que este algoritmo utiliza los recursos de EuroWordNet, y dado que esta base aún se encuentra en desarrollo, se usó el reagrupamiento de palabras por familia. De esta forma, en lugar de normalizar cada palabra por su lema, estas se sustituyen por un representante de su

familia léxica. La probabilidad de localizar una palabra en EuroWordNet aumenta al utilizar el representante de la familia léxica en lugar de su lema.

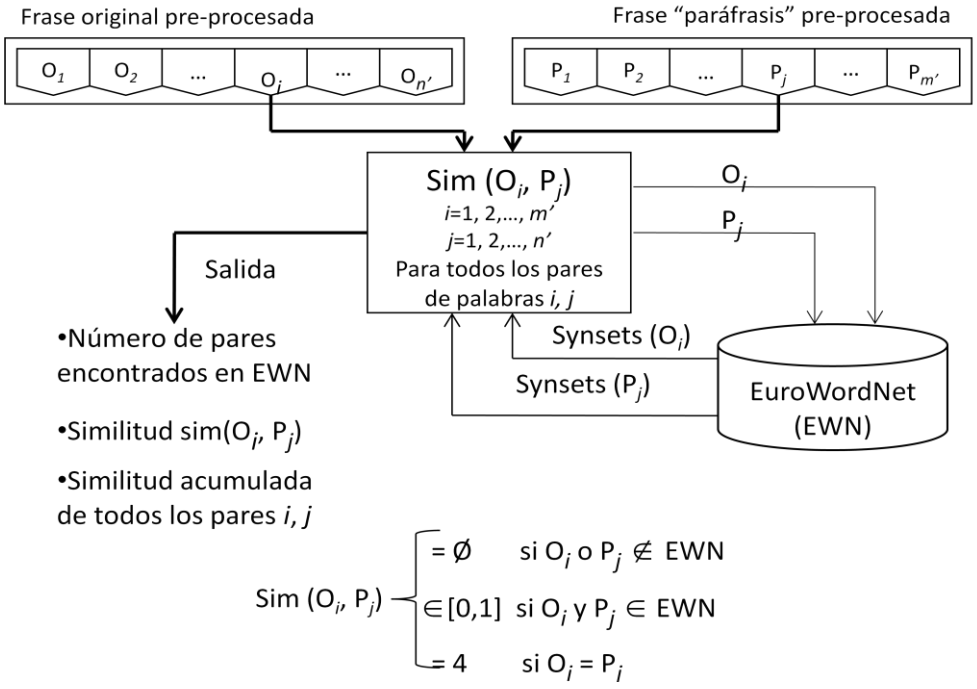


Figura 5.5 Arquitectura de sim_rst

La Figura 5.5 muestra un esquema de la arquitectura del segundo paso de este algoritmo. Este paso se incluye en la arquitectura general como "sim_rst" y funciona como sigue: cada una de las palabras $O_{m'}$ del fragmento de texto $F_{o'}$ se compara con cada una de las palabras $P_{n'}$ del fragmento de texto $F_{p'}$. $F_{o'}$ está formado por *m'* número de palabras y $F_{p'}$ está formado por *n'* número de palabras. Para hacer esta comparación y dado que *n'* no necesariamente es igual a *m'*, se comparan *n'* x *m'* parejas.

Primeramente se hace una búsqueda en EuroWordnet de los términos O_i y P_j a comparar. Sólo se comparan verbos con verbos y sustantivos con sustantivos, dada la arquitectura actual de EuroWordNet. Tal como hicieron Vivaldi *et al.* (2010), "we use

information obtained from the hyperonimical paths for each *synset* (sy) in EuroWordNet."[32] Con esta información se usa la fórmula siguiente para calcular la similitud entre palabras:

$$Sim(sy_1, sy_2) = \frac{2 \times \#Common\,Nodes(sy_1, sy_2)}{Depth(sy_1) + Depth(sy_2)}$$

Figura 5.6 Fórmula para calcular la similitud semántica entre pares de palabras

Este cálculo se traduce como: la similitud entre dos *synsets* es igual al número de nodos comunes de los *synsets* multiplicado por 2 y dividido entre la profundidad de uno de los *synsets* más la profundidad del otro. Vivaldi *et al*. (2010) proporcionan un ejemplo de este cálculo usando las palabras en inglés "vas" y "gland":

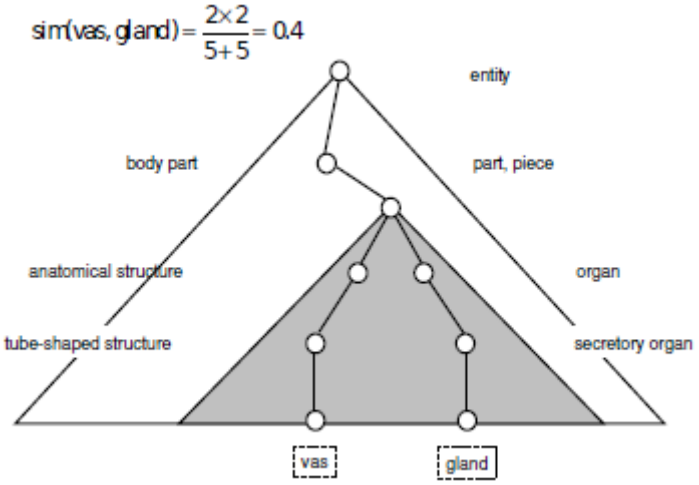


Figura 5.7 Ejemplo de cálculo de similitud semántica

Como se ve, contando desde la cima de la pirámide, ambas palabras comparten 2 nodos, dado que el nodo en el que ambas palabras se separan no se cuenta como común. También se observa en la tabla que la profundidad de ambos términos es de 5, de igual manera contando desde la cima de la pirámide y tomando en cuenta que los nodos en la base de la pirámide no se cuentan para la profundidad, dado que representan la palabra en sí misma. Entonces, siguiendo la fórmula para este ejemplo, se multiplica 2 por 2 y se divide

[32] "Usamos la información obtenida de las rutas hiperonímicas de cada *synset* (sy) en EuroWordNet."

entre 5 más 5, lo que nos da como resultado la similitud semántica entre estos dos términos, o sea 0.4.

El algoritmo utilizado en este trabajo, continúa de la siguiente manera: se toma un término O_i del fragmento de texto original u $F_{o'}$ y un término P_j del texto con el que se compara el original o $F_{p'}$, y se buscan en EuroWordNet. Se calcula su similitud mediante la fórmula de la Figura 5.6. Para ello, si alguna o las dos palabras no se encuentran en EuroWordNet, el resultado de su similitud es 0. Si las dos palabras a comparar son idénticas, el valor de su similitud será de 4. Este último valor se definió dado que, en primer lugar, la utilización repetida de palabras idénticas es una marca de alto grado de paráfrasis y, en segundo lugar, empíricamente encontramos que este valor compensa la gran cantidad resultante de valores 0 de similitud. Esta gran cantidad de similitudes 0 se debe en su mayoría a las carencias de EuroWordNet. Si no se diera alguno de los dos casos anteriores, la similitud se calcula con la fórmula de la Figura 5.6, y será un valor normalizado entre 0 y 1.

Este cálculo de similitud se hace entre todos los pares posibles de palabras, por lo que se realiza *n'* x *m'* veces. Después de cada cálculo, se acumula el resultado para obtener un valor único para la pareja de textos $F_{o'}$ y $F_{p'.}$ Posteriormente se divide entre el número total de parejas de palabras para obtener un resultado final dividiendo el valor acumulado entre *n'* x *m'*. Ese resultado final representa el valor de la similitud semántica entre los fragmentos de texto $F_{o'}$ y $F_{p'}$ y se muestra como una cantidad entre 0 y 1, en donde 0 significa completa ausencia de similitud y 1 significa similitud total.

Por ejemplo, al comparar las EDUs a y b se obtuvo una similitud de 0.714. Este resultado, dado que es una cantidad muy cercana a 1, indica que estas dos frases son muy

similares. O sea, si tomamos en cuenta que un resultado de 1 implica un 100% de similitud semántica, estos dos fragmentos son semánticamente similares en un 71.4%, lo que es alto.

a) [aunque sí la logran y la mayor parte de las veces muy intensa, por la estimulación adecuada del clítoris.] SATÉLITE_Concesión
b) [pero si lo consiguen, y de forma muy intensa, por medio de la estimulación correcta del clítoris.] SATÉLITE_Concesión

Por otra parte, al hacer este cálculo de similitud semántica entre las EDUs c y d resultó una similitud de 0.067. Puesto que este resultado se acerca mucho a cero, nos indica que estas frases son muy diferentes. En otros términos podemos decir que estas frases son semánticamente similares en un 6.7%, lo cual, tomando en cuenta que ambas frases provienen de textos del mismo ámbito y temática, es bajo.

c) [de lo anterior desprende el hecho de que muchas mujeres no logran la satisfacción orgásmica por la penetración] SATÉLITE_Resultado
d) [La información abrumadora de la disfunción sexual masculina no permite ver su contraparte femenina.] SATÉLITE_Resultado

5.4.2. Verificación de nuestros resultados mediante cálculo de similitud semántica

Siguiendo el algoritmo descrito en el punto 5.4.1, se aplicó el programa creado por el Dr. Torres-Moreno a las EDUs que según el análisis que se describe en el apartado 5.3 coincidían en tanto a relación discursiva y contenido. Se llevaron a cabo estas comparaciones separando núcleos de satélites, por lo que para la comparación de tipo A se calculó la similitud semántica entre 16 pares de EDUs, para la comparación de tipo B se

calculó entre 28 pares de EDUs y para la comparación de tipo C se hizo entre 18 pares de EDUs.

Tabla 5.7 Resultados similitud semántica

	Similitud semántica	Soporte CORTEX	% soporte CORTEX	Número de parejas
Promedio A	0.317	10.38	20.28%	54.81
Promedio B	0.428	14.82	19.59%	91.43
Promedio C	0.070	10.67	10.81%	90.78

Con el promedio de los resultados que arroja el programa en perl hecho por Juan-Manuel Torres-Moreno se formó la Tabla 5.7. La primera columna de la tabla enuncia la comparación cuyos resultados se observan en el resto de las columnas. Por su parte, la segunda columna de esta tabla muestra el promedio de las similitudes semánticas obtenidas para las comparaciones A, B y C.

La tercera columna muestra el promedio de palabras encontradas por el agrupador de familias léxicas de CORTEX, también llamado soporte. La cuarta columna contiene el porcentaje de términos encontrados por el agrupador de familias léxicas de CORTEX. Y, finalmente, la quinta columna muestra el promedio de cuántas parejas de palabras se compararon. Recordemos que estas parejas indican el número de veces, en promedio, que se hizo el cálculo de similitud textual. En otras palabras, esa columna presenta el número de veces que se compararon pares de palabras por cada tipo de comparación. Es importante tomar en cuenta que son promedios y por esto tenemos valores decimales aún si se trata de un conteo de palabras.

Como podemos observar, el promedio de las similitudes semánticas obtenidas en las comparaciones de tipo A y B son de 0.317 y 0.428 respectivamente, mientras que para la

comparación de tipo C se obtuvo un promedio de 0.070. Si dividimos el promedio de las similitudes semánticas obtenidas en las comparaciones de tipo A entre el promedio de las comparaciones de tipo C, obtenemos la razón aritmética entre la similitud textual de A y C. El resultado de 4.50 significa que, según nuestros resultados, las EDUs entre las que se midió la similitud semántica para la comparación de tipo A son cuatro veces y media más parecidas entre sí que las que se utilizaron para medir la similitud semántica en la comparación de tipo C. A su vez, si se calcula la razón entre las similitudes semánticas obtenidas de los corpus B y C, se obtiene un resultado de 6.08. Esto nos dice que, según nuestroanálisis, las EDUs entre las que se midió la similitud semántica para la comparación de tipo B son alrededor de seis veces más parecidas entre sí que las que se utilizaron para medir la similitud semántica en la comparación de tipo C.

En la Figura 5.8 se presenta una gráfica de barras que muestra los resultados según el nivel de paráfrasis que había en las comparaciones. Cada columna de esta gráfica representa el valor promedio de similitud semántica que se obtuvo en las comparaciones. La columna etiquetada como “Baja” se refiere a la similitud semántica promedio para la comparación B, que como recordaremos es la comparación entre los textos originales y las paráfrasis de nivel bajo. La siguiente columna, etiquetada como “Alta”, representa la similitud semántica promedio encontrada entre los textos originales y las paráfrasis de nivel alto, o sea la comparación A. La última columna hace referencia al promedio de similitud semántica encontrada entre los textos originales y los textos no-paráfrasis o comparación C.

En esta gráfica podemos observar que en la comparación B, que contenía un nivel bajo de paráfrasis, se encontró un nivel más alto de similitud con un valor de 0.428. Luego, la comparación A, que se hizo con un nivel alto de paráfrasis, tiene un valor medio con

0.317. Finalmente la similitud más baja, como era de esperarse, se muestra en la comparación C que contenía textos en los que no había paráfrasis.

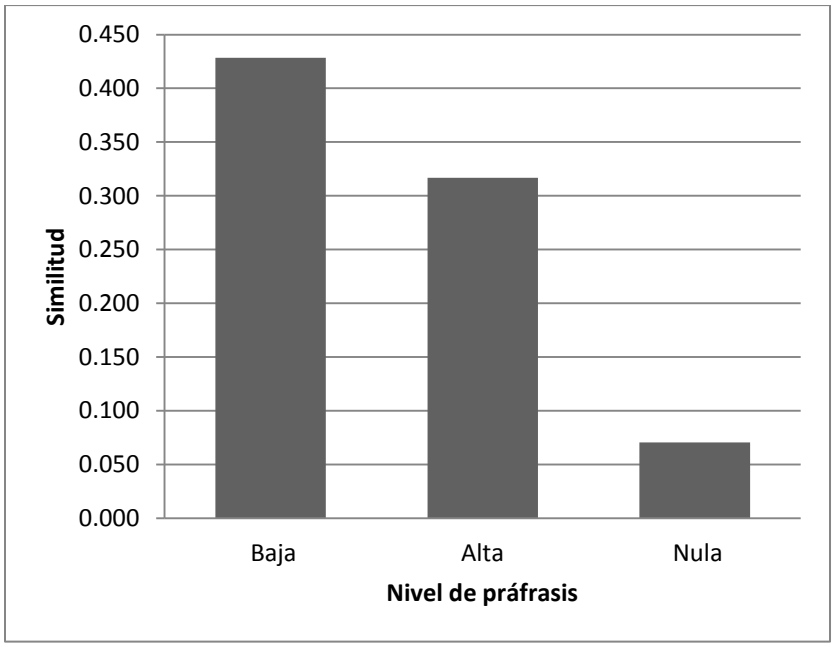


Figura 5.8 Gráfico de resultados de similitud semántica

Como era de esperarse, nuestros resultados muestran una similitud más alta entre los textos originales OR y sus paráfrasis Pa y Pb que entre los textos originales OR y los que no eran paráfrasis o Pno. A pesar de esto, dado que el porcentaje de soporte que obtuvimos es muy bajo para obtener resultados definitivos es necesario que se vea la manera de aumentar el soporte.

5.5. Discusión de los resultados

Con base en los resultados obtenidos, podemos ver que es posible distinguir los casos en los que nos encontramos con paráfrasis de textos. Además, atendiendo a la cantidad de coincidencias encontradas en el análisis manual (ver Tabla 5.6) y al cálculo de similitud semántica (ver Figura 5.8), se puede identificar el nivel de complejidad de una paráfrasis.

Podríamos decir que para llevar a cabo un dictamen de detección de similitud textual, la parte manual de nuestro método acepta o niega la existencia de una paráfrasis dado un par de textos a comparar. Por su parte, el análisis automático que llevamos a cabo confirma lo dado por el análisis manual. Además, si tomamos nuestros resultados de similitud semántica (ver Figura 5.8) como umbrales para los niveles de paráfrasis, también es posible establecer el nivel de complejidad que presenta una paráfrasis.

Dados los valores que obtuvimos para los niveles de paráfrasis, incluso podría pensarse que pueden ser usados como base para un clasificador automático de textos por nivel de paráfrasis.

Finalmente, si observamos las figuras Figura 5.2 y Figura 5.3, podríamos pensar que si se tomaran en cuenta para el análisis las relaciones multinucleares y de Elaboración probablemente aumentaría la cantidad de coincidencias entre los textos.

6.Conclusiones

En este capítulo se presentan las conclusiones que derivan de esta tesis. Primero se describirán con referencia a las hipótesis y objetivos de este trabajo y finalmente se presentarán las posibles aplicaciones relacionadas con los resultados de este trabajo.

6.1. Conclusiones referentes a las hipótesis

Antes de describir las conclusiones respecto a las hipótesis que se plantean en este trabajo, recordemos éstas como se presentan en el capítulo introductorio.

Hipótesis primarias:

- Las estructuras discursivas de un texto original y un texto que lo parafrasee han de ser similares, aunque el léxico y la estructura sintáctica sean diferentes.
- Ha de ser posible establecer la similitud entre el contenido semántico de un texto original y el de un texto que lo parafrasee mediante un cálculo.

Hipótesis secundarias:

- Con la comparación de las estructuras discursivas de dos textos ha de ser posible identificar las unidades discursivas que poseen similitudes entre un texto y otro.
- A pesar de que las estructuras de dos textos pertenecientes al mismo género discursivo pudieran ser parecidas, un texto que parafrasea a otro (tomado como original) presentará mayor parecido con éste que la estructura de otro texto que no representa una paráfrasis con el mismo original.

- Un texto que parafrasee a otro habrá de presentar mayor similitud semántica con éste que un texto que no lo haga.

Teniendo estas hipótesis en mente, a continuación se describen las conclusiones que derivan de ellas separándolas según si fueron planteadas como hipótesis primarias o secundarias.

Validación de las hipótesis primarias

Respecto a la primera hipótesis podemos decir que nuestros resultados apuntan a su validez. Es así que podemos ver que entre los textos analizados presentados como no paráfrasis y los textos tomados como originales, se observó una cantidad menor de relaciones discursivas coincidentes. Debe tomarse en cuenta que no sólo se comparó el tipo de relación entre cada par de EDUs, sino que, para constatar esta hipótesis, se verificó el contenido semántico de éstas.

Tomando en consideración la segunda hipótesis se pudo constatar que, en efecto, es posible calcular la similitud semántica entre textos. Este cálculo, hay que recordar, se hizo entre las EDUs que, dada la comparación de estructuras discursivas, se vio que coincidían. Y no sólo es posible, sino que el resultado del cálculo de similitud semántica es distinto si se trata de una comparación entre textos originales y sus paráfrasis que si se trata de la comparación entre textos originales y otros de la misma temática pero que no sean paráfrasis (ver Figura 5.8).

Validación de las hipótesis secundarias

De la primera hipótesis secundaria podemos decir que los resultados obtenidos apuntan a su validez. Como primera comprobación de esta hipótesis, mediante el cálculo de similitud semántica llevado a cabo en la metodología de esta tesis se verificó la similitud semántica

entre las unidades discursivas cuyas relaciones discursivas se observaron similares. Con este cálculo se pudo observar que en las EDUs coincidentes, según la comparación de estructuras discursivas, se presentó un alto nivel de similitud semántica. Lo anterior nos indica que es posible que la comparación de estructuras discursivas ayude a la distinción de las EDUs cuyo contenido es similar entre un texto y otro. Además, como segunda verificación de la similitud entre las unidades discursivas cuyas relaciones coincidían, se calculó la similitud semántica entre algunas EDUs obtenidas de los textos originales y los que no eran paráfrasis, las cuales según la comparación de estructuras discursivas no eran coincidentes. De esta última comparación también se obtuvieron resultados cercanos a lo que se esperaba.

Respecto a la segunda hipótesis de este tipo, podemos ver que los resultados de las tablas, como la Tabla 5.4, cuya construcción se describe en el apartado 5.3.2 de la presente tesis, apuntan a su validación. Dichos resultados muestran que una cantidad mayor de EDUs y sus relaciones resultan similares, dada una comparación de las estructuras retóricas, cuando se comparan los originales con los textos paráfrasis que cuando se comparan con los textos no paráfrasis. Digamos, hay más coincidencias en la estructura retórica de un texto original y su paráfrasis que entre un texto original y otro de la misma temática que no sea paráfrasis. Es ahí que podemos observar que, a pesar de que la diferencia numérica no se observa muy grande en las tablas descritas en el mencionado apartado 5.3.2, sí hay una diferencia cuantitativa entre las coincidencias encontradas en ambas comparaciones, lo que nos lleva a pensar que se podrían encontrar datos más contundentes en análisis futuros basados en la metodología del presente trabajo.

La validez de la tercera hipótesis de este tipo fue confirmada por los resultados obtenidos de la metodología presentada en esta tesis. En la Figura 5.4 es posible observar

que efectivamente hay una diferencia cuantitativa observable entre las comparaciones de los textos originales y sus respectivas paráfrasis, dependiendo si son de “nivel alto” o “nivel bajo”.

6.2. Conclusiones referentes a los objetivos

Ahora que ya se ha hecho la validación de las hipótesis, es posible concluir respecto a los objetivos planteados en el primer capítulo de esta tesis. A continuación se hará la mención de cada objetivo y se describirá si fue alcanzado o no.

- Comparar las estructuras discursivas de un corpus de textos originales y de textos parafraseados a distintos niveles, para observar si estas estructuras discursivas son similares.

Este primer objetivo fue alcanzado, ya que se obtuvieron tablas como las mencionadas en el apartado 5.3 de la presente tesis. Gracias a estas tablas fue posible observar si las estructuras discursivas de los textos originales y los textos parafraseados del corpus utilizado en esta tesis son similares. Gracias a esto, se pudo hacer la comprobación de varias de las hipótesis planteadas en el capítulo 1 de este trabajo.

- Calcular la similitud semántica entre los textos originales y los textos parafraseados a distintos niveles pertenecientes a un corpus, para observar si el resultado obtenido es útil para la detección de similitud textual.

Este segundo objetivo también se consiguió según lo mencionado en el apartado 5.4 de este trabajo. A su vez, esto se puede ver en que los resultados descritos en el apartado mencionado no sólo resultaron útiles para la detección de similitud textual, sino que además resultan distinguibles los niveles de paráfrasis que se tienen en el corpus aquí utilizado.

- Comprobar si la comparación de las estructuras discursivas de textos originales y textos que los parafraseen, junto con el cálculo de su similitud semántica pueden resultar de utilidad para la detección de similitud textual.

Este último objetivo se logró gracias a que, en un análisis de los resultados obtenidos de la metodología propuesta en este trabajo, se observan resultados cuantitativos de similitud entre los textos que forman el corpus utilizado.

6.3. Contribuciones de este trabajo

La principal aportación de este trabajo deriva del hecho que, hasta donde esta autora tiene conocimiento, no hay metodologías que se basen tanto en análisis discursivo como en cálculos semánticos para la detección de similitud textual. Aquí se presenta un método que sí lo hace, por lo que es posible tomar como innovador el método aquí propuesto.

Además, derivados de este trabajo se encuentran al menos dos artículos de pendiente publicación. Uno para el XI Congreso Nacional de Lingüística con realización en Chetumal, Quintana Roo y otro dentro del III Workshop "A RST e os Estudos do Texto" que se llevará a cabo en Cuiabá, Brasil.

6.4. Posibles aplicaciones

Teniendo en cuenta los resultados obtenidos en este trabajo, es posible observar que, además de que la detección de similitud textual se puede utilizar en la detección de plagio, esta metodología se puede utilizar para varias aplicaciones. En primer lugar, puede utilizarse para la evaluación de exámenes, si se compararan las respuestas de los exámenes de las personas examinadas con un texto previamente resuelto por el examinador. Si hubiera similitud alta, las respuestas serían presumiblemente correctas.

En segundo lugar, puesto que la RST es, hasta cierto punto, independiente de la lengua y dado que EuroWordNet existe en varios idiomas, sería posible implementar esta metodología para más idiomas.

6.5. Trabajo futuro

Puesto que los resultados obtenidos en este trabajo apuntan a la comprobación de las hipótesis planteadas en éste, mas no las comprueban por completo, se puede concluir que es necesaria la aplicación futura de una metodología basada en el trabajo descrito en la presente tesis para verificar que los resultados obtenidos no fueran producto de una casualidad.

Es así también que, dada la experiencia empírica obtenida mediante la realización del trabajo aquí presentado, se observó que en un trabajo futuro probablemente se pudiera llevar a cabo esta metodología sin la necesidad de realizar el paso descrito en el apartado 5.3.2. ya que para el análisis automático pudiera no ser necesaria una comparación semántica manual previa.

También, sería interesante aplicar el método presentado en esta tesis tomando también en cuenta las relaciones multinucleares y de Elaboración. Ya que esta tesis era una primera aproximación estas relaciones discursivas no se tomaron en cuenta, pero en una nueva aplicación del método aquí descrito sería interesante ver los resultados obtenidos con la inclusión de estas relaciones.

Además, para una nueva verificación de los resultados obtenidos, es necesaria la incrementación del tamaño del corpus utilizado. Esta ampliación del corpus se podría llevar a cabo tomando el RST Spanish Treebank[33] como fuente de documentos originales. Esta

[33] http://corpus.iingen.unam.mx/rst/

fuente de documentos originales presentaría la ventaja de que los textos ya se encuentran analizados discursivamente mediante la RST. Posteriormente, también respecto al RST Spanish Treebank, se podría incluir el corpus de paráfrasis para que la comunidad científica que trabaja en detección de similitud textual en español pueda hacer uso de él.

6.6. Conclusiones generales

Finalmente, la metodología propuesta en este trabajo no sólo se mostró útil en la detección de similitud textual, sino que también, gracias a los resultados que presentan, se observa la posibilidad de distinguir entre niveles de paráfrasis. Es decir, gracias a este método no sólo se puede definir si hay o no similitud entre pares de textos, sino que se puede decir qué nivel de complejidad presenta una paráfrasis.

Bibliografía

Arano, S. (2003). "*La ontología: una zona de interacción entre la Lingüística y la Documentación"* [en línea]. *Hipertext.net*, núm. 2. Barcelona, España: **Universitat Pompeu Fabra.** Recuperado el 15 de junio de 2011, de: http://www.hipertext.net

Barrón-Cedeño, A., Vila, M., y Rosso, P. (2010). "Detección automática de plagio: de la copia exacta a la paráfrasis" [Versión electrónica]. En *Panorama actual de la lingüística forense en el ámbito legal y policial: Teoría y práctica (Jornadas (in)formativas de lingüística forense).* Madrid: Euphonia Ediciones SL. 76-96. Recuperado el 17 de junio de 2011, de: http://users.dsic.upv.es/~lbarron/publications/2010/BarronEtAl_JLF10-1.pdf

Brandt, J., Gutbrod, M., Wellnitz, O., y Wolf, L. (2010). "Plagiarism Detection in Open Access Publications" [Versión electrónica]. En actas de la *4th International Plagiarism Conference.* Recuperado el 23 de junio de 2011, de: *http://www.plagiarismadvice.org/documents/conference2010/papers/4IPC_0029_final.pdf*

Calsamiglia Blancafort, H., y Tusón Valls, A. (2007). *Las cosas del decir: Manual de análisis del discurso.* (2ª ed.) Barcelona: Ariel. 363 p.

Campos Plaza, N., y Ortega Arjonilla, E. (2005). *Panorama de lingüística y traductología: aplicaciones a los ámbitos de la enseñanza del francés/lengua extranjera y de la traducción (francés-español).* La Mancha: Universidad de Castilla. 832 p.

Carlson, L. and Marcu, D. (2001). *Discourse Tagging Reference Manual* [Reporte técnico]. Information Sciences Institute. ISI-TR-545. Recuperado el 25 de junio de 2011, de: http://www.isi.edu/~marcu/discourse

Clough, P. (2000). *Plagiarism in natural and programming languages: an overview of current tools and technologies* [Reporte técnico]. University of Sheffield, Dept. of computer Science, Reino Unido. CS-00-05. Recuperado el 15 de junio de 2011, de: ftp://ftp.dlsi.ua.es/people/armando/maria/Plagiarism.rtf

Clough, P. (2003). *Old and new challenges in automatic plagiarism detection* [documento electrónico]. Newcastle-uponTyne: JISC National Plagiarism Advisory Service. Recuperado el 17 de junio de 2011, de: http://ir.shef.ac.uk/cloughie/index.html

da Cunha Fanego, I., Torres-Moreno, J.-M., y Sierra Martínez, G. (2011). "On the Development of the RST Spanish Treebank". En *Proceedings of the Fifth Law Workshop (LAW V)*. Portland, Oregon. Association for Computational Linguistics. 1-10. Recuperado el 20 de junio de 2011, de: http://aclweb.org/anthology-new/W/W11/W11-0401.pdf

da Cunha, I., E. SanJuan, J.-M. Torres-Moreno, M. Lloberes, and I. Castellon. (2010). "DiSeg: Un segmentador discursivo automatico para el español" [Versión

electrónica]. *Procesamiento de Lenguaje Natural*, Vol. 45. ISSN: 1989-7553. http://ararat.ujaen.es/sepln/ojs/ojs-2.3.5/index.php/pln/article/view/776/630

da Cunha, I., y Iruskieta, M. (2010). "Comparing rhetorical structures of different languages: The influence of translation strategies". *Discourse Studies, 12* (5), 563-598.

Fellbaum, C. (1990). "English verbs as a semantic net". *International Journal of Lexicography, 3* (4), 40-61.

Fellbaum, C. (Ed.) (1999). *WordNet: An electronic Lexical Database.* Massachusetts: MIT Press. 423 p. ISBN-13 9780262061971

Fellbaum, C., Gross, D., y Miller, K. (1990). "Adjectives in WordNet". *International Journal of Lexicography, 3* (4), 26-39.

Frías Conde, X. (2002). **"**Introducción a la psicolingüística" [en linea]. *Ianua: Revista Philologica Romanica*. Sup. 06. ISSN: 1616-413X. Recuperado el 29 de junio de 2011, de: http://www.romaniaminor.net/ianua/sup/sup07.pdf

Kang, N., Gelbukh, A., y Han, S. (2006). "PPChecker: Plagiarism Pattern Checker in Document Copy". En *Lecture Notes in Computer Science* 4188. 661-667.

Lope Blanch, J. M. (1987). *Análisis gramatical del discurso*. México: Universidad Nacional Autónoma de Mexico. 254 p. ISBN 968-36-0230-4

López Ferrero, C. (2002). "Aproximación al análisis de los discursos profesionales" [en linea]. *Revista signos* ISSN: 0718-0934. Recuperado el 30 de junio de 2011, de: http://www.scielo.cl/scielo.php?pid=S0718-93420020051000013&script=sci_arttext

Maingueneau, D. (1980). *Introducción a los métodos de análisis del discurso: Problemas y perspectivas.* (L. Castro, Trans.) Buenos Aires, Argentina: Librería Hachette. 212 p.

Mann, W. C., y Thompson, S. A. (1988). "Rhetorical Structure Theory: Toward a functional theory of text organization". *Text, 8* (3), 244-281.

Maurer, H., Kappe, F., y Zaka, B. (2006). "Plagiarism - A Survey". *Journal of Universal Computer Science, 12* (8), 1050- 1084.

Miller, G. (1990). "Nouns in WordNet: A Lexical Inheritance System". *International Journal of Lexicography , 3* (4), 10-25.

Miller, G. A. (1995). "WordNet: A Lexical Database for English". *Communications of the ACM*, 38 (11), 39-41.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller. (1993). "Introduction to WordNet: An On-line Lexical Database" [en línea]. *International Journal of Lexicography,* 3(4):235-244. DOI: 10.1093/ijl/3.4.235

Moreno Quibén, N. (2008). *Semántica Léxica* [Apuntes de la materia comprensión y producción]. España: Centro de Estudios Universitarios de la Universidad de Castilla - La Mancha. Recuperado el 26 de junio de 2011, de: http://www.quiben.org/32030

O'Donnell, M. (2000). "RSTTOOL 2.4 – A markup tool for rhetorical structure theory". En *Proceedings of the International Natural Language Generation Conference*. 253-256.

Portoles, J. (2002) *Marcadores del discurso*. (4ª ed.) Barcelona: Ariel. 183 p. ISBN 8434482460

Princeton University. (2010). *About WordNet*. Princeton University. Recuperado el 15 de junio de 2011, de: http://wordnet.princeton.edu

Renkema, J. (1999). *Introducción a los estudios del discurso*. Barcelona: Gedisa. 285 p.

Rosas González, A. (2011). *Análisis estilométrico para la detección de plagio.* Tesis de licenciatura, Universidad Nacional Autónoma de México, Ciudad de México, México.

Sánchez Vega, J. F. (2011). *Detección automática de plagio basada en la distinción y fragmentación del texto reutilizado.* Tesis de maestría, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Puebla, México.

Shivakumar, N., y García-Molina, H. (1995). "SCAM: A Copy Detection Mechanism for Digital Documents". [Versión digital]. *Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries.* Austin, Texas. Recuperado el 20 de junio de 2011, de: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.6880yrep=rep1ytype=pdf

Si, A., Leong, H., y Lau, R. (1997). "CHECK: A Document Plagiarism Detection System". *Proceedings of ACM Symposium for Applied Computing*, 70-77.

Torres-Moreno, J.-M., Velázquez, P., y Meunier, J.-G. (2001). "Cortex: un algorithme pour la condensation automatique des texts". *La cognition entre individu et société : modèles et méthodes*, Actas de *ARCo 2001*. Lyon,France. 365 p. ISBN 2-746203588

Torres-Moreno, J.-M., Velázquez, P., y Meunier, J.-G. (2002). "Condensés de textes par des méthodes numériques". *6th International Conference on the Statistical Analysis of Textual Data (JADT 2002): 6es Journées Internationales d'Analyse statistique des Données Textuelles*. Versión digital disponible en http://lexicometrica.univ-paris3.fr/jadt/jadt2002/PDF-2002/torres_velazquez_meunier.pdf

Van Dijk, T.A. (2000). El estudio del discurso. En T. A. Van Dijk (Comp.). *El discurso como estructura y proceso* (pp.21-65). Barcelona: Gedisa.

Vivaldi, J., da Cunha, I., Torres-Moreno, J.-M., y Velázquez, P. (2010). "Automatic Summarization Using Terminological and Semantic Resources". *7th International Conference on Languaje Resources and Evaluation (LREC 2010).* Valleta, Malta.

Collins English Dictionary (2000). *The Collins English Dictionary*. HarperCollins Publishers.

Anexos

Anexo1. Tamaño del sub-corpus de temática sushi

Sushi			
Texto	**Palabras**	**EDUs**	**Relaciones**
OR	791	54	47
Pa	630	51	48
Pb	598	46	39
NoP	494	32	31
Total	**2513**	**183**	**165**

Anexo2. Tamaño del sub-corpus de temática sexualidad

Sexualidad			
Texto	**Palabras**	**EDUs**	**Relaciones**
OR	704	33	33
Pa	518	29	28
Pb	572	36	34
NoP	642	41	35
Total	**2436**	**139**	**130**

Anexo3. Tamaño del sub-corpus de temática astronomía

Astronomía			
Texto	**Palabras**	**EDUs**	**Relaciones**
OR	929	40	33
Pa	1026	50	37
Pb	923	40	33
NoP	740	31	30
Total	**3618**	**161**	**133**

Anexo4. Comparación de estadísticas sub-corpus de temática sushi

Comparación A sushi				
	Original Sushi		Paráfrasis alta sushi	
	Relación	Cantidad	Relación	Cantidad
Relaciones Multinucleares	Contraste	1	Contraste	2
	Secuencia	3	Secuencia	2
	Unión	4	Unión	5
Relación de Elaboración	Elaboración	18	Elaboración	16
Relaciones Únicas	Circunstancia	1		
			Medio	1
	Lista	1		
			Resumen	1
Relaciones Coincidentes	Antítesis	2	Antítesis	1
	Causa	4	Causa	3
	Concesión	1	Concesión	2
	Fondo	4	Fondo	4
	Motivación	1	Motivación	1
	Propósito	3	Propósito	5
	Resultado	3	Resultado	5

Comparación B sushi				
	Original sushi		Paráfrasis baja sushi	
	Relación	Cantidad	Relación	Cantidad
Relaciones Multinucleares	Contraste	1	Contraste	1
	Lista	1	Lista	1
	Secuencia	3	Secuencia	3
	Unión	4	Unión	4
Relación de Elaboración	Elaboración	18	Elaboración	15
Relaciones Únicas	Motivación	1		
Relaciones Coincidentes	Antítesis	2	Antítesis	1
	Causa	4	Causa	5
	Circunstancia	1	Circunstancia	1
	Concesión	1	Concesión	1
	Fondo	4	Fondo	4
	Propósito	3	Propósito	1
	Resultado	3	Resultado	2

Comparación C sushi				
	Original sushi		No paráfrasis sushi	
	Relación	Cantidad	Relación	Cantidad
Relaciones Multinucleares	Secuencia	3	Secuencia	3
	Unión	4	Unión	2
Relación de Elaboración	Elaboración	18	Elaboración	12
Relaciones Únicas	Antítesis	2	Disyunción	1
	Contraste	1		
	Lista	1		
	Motivación	1		
Relaciones Coincidentes	Causa	4	Causa	2
	Circunstancia	1	Circunstancia	1
	Concesión	1	Concesión	1
	Fondo	4	Fondo	7
	Propósito	3	Propósito	1
	Resultado	3	Resultado	2

Anexo5. Comparación de estadísticas sub-corpus de temática sexualidad

Comparación A sexualidad				
	Original Sexualidad		Paráfrasis alta sexualidad	
	Relación	Cantidad	Relación	Cantidad
Relaciones Multinucleares	Contraste	1	Contraste	1
	Disyunción	1		
	Lista	1		
	Secuencia	1	Secuencia	1
	Unión	3	Unión	2
Relación de Elaboración	Elaboración	16	Elaboración	13
Relaciones Únicas	Causa	1	Circunstancia	1
			Conjunción	1
			Evaluación	1
			Medio	1
Relaciones Coincidentes	Concesión	3	Concesión	1

	Fondo	1	Fondo	1
	Interpretación	3	Interpretación	4
	Resultado	2	Resultado	1

Comparación B sexualidad				
	Original Sexualidad		**Paráfrasis baja sexualidad**	
	Relación	Cantidad	Relación	Cantidad
Relaciones Multinucleares	Contraste	1	Contraste	2
	Disyunción	1		
	Lista	1	Lista	1
	Secuencia	1		
	Unión	3	Unión	3
Relación de Elaboración	Elaboración	16	Elaboración	12
Relaciones Únicas			Circunstancia	3
			Condición	1
			Conjunción	1
Relaciones Coincidentes	Causa	1	Causa	1
	Concesión	3	Concesión	3
	Fondo	1	Fondo	2
	Interpretación	3	Interpretación	3
	Resultado	2	Resultado	2

Comparación C sexualidad				
	Original Sexualidad		**No paráfrasis sexualidad**	
	Relación	Cantidad	Relación	Cantidad
Relaciones Multinucleares	Contraste	1		
	Disyunción	1		
	Lista	1	Lista	2
	Secuencia	1		
	Unión	3	Unión	5
Relación de Elaboración	Elaboración	16	Elaboración	14
Relaciones Únicas	Causa	1	Antítesis	2

	Concesión	3		
			Evaluación	1
			Evidencia	2
	Interpretación	3		
			Motivación	1
			Propósito	2
Relaciones Coincidentes	Fondo	1	Fondo	1
	Resultado	2	Resultado	5

Anexo6. Comparación de estadísticas sub-corpus de temática astronomía

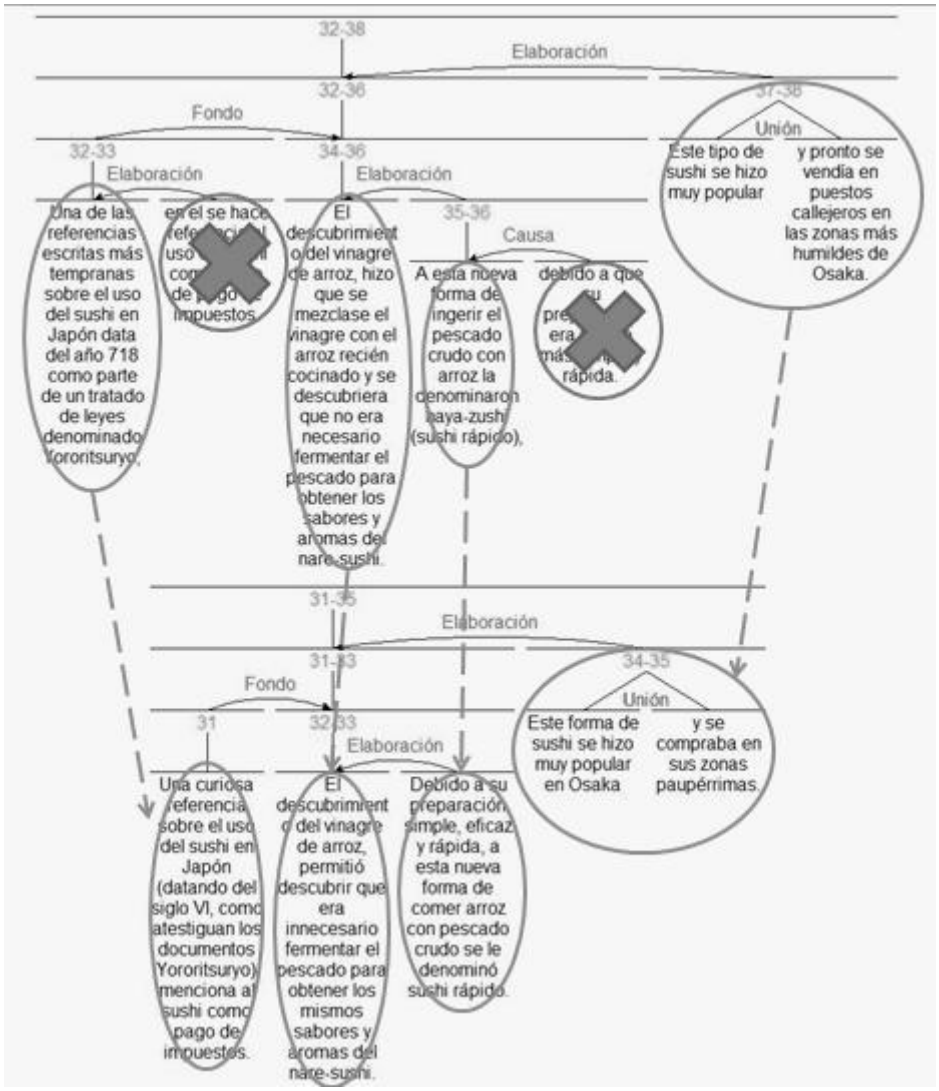
Comparación A astronomía				
	Original astronomía		**Paráfrasis alta astronomía**	
	Relación	Cantidad	Relación	Cantidad
Relaciones Multinucleares	Contraste	1		
	Lista	4	Lista	4
	Secuencia	3	Secuencia	3
	Unión	5	Unión	5
Relación de Elaboración	Elaboración	15	Elaboración	16
Relaciones Únicas			Antítesis	1
			Causa	1
			Circunstancia	2
Relaciones Coincidentes	Evidencia	1	Evidencia	1
	Medio	1	Medio	1
	Propósito	1	Propósito	2
	Resultado	2	Resultado	1

Comparación B astronomía				
	Original astronomía		**Paráfrasis baja astronomía**	
	Relación	Cantidad	Relación	Cantidad
Relaciones Multinucleares	Contraste	1	Contraste	1
	Lista	4	Lista	4
	Secuencia	3	Secuencia	3
	Unión	5	Unión	6
Relación de Elaboración	Elaboración	15	Elaboración	15
Relaciones Coincidentes	Evidencia	1	Evidencia	1
	Medio	1	Medio	1

	Propósito	1	Propósito	1
	Resultado	2	Resultado	1

Comparación C astronomía				
	Original astronomía		**No paráfrasis astronomía**	
	Relación	Cantidad	Relación	Cantidad
Relaciones Multinucleares	Contraste	1	Contraste	1
	Lista	4	Lista	2
	Secuencia	3	Secuencia	1
	Unión	5	Unión	1
Relación de Elaboración	Elaboración	15	Elaboración	13
Relaciones Únicas	Evidencia	1		
			Fondo	3
			Justificación	1
			Motivación	1
			Preparación	2
Relaciones Coincidentes	Medio	1	Medio	1
	Propósito	1	Propósito	2
	Resultado	2	Resultado	2

Anexo7. Ejemplo de comparación visual entre fragmentos de estructuras discursivas etiquetadas con la herramienta RSTtool



Anexo8. Tabla de resultados del cálculo de similitud semántica en el corpus total

Resultados sim_rst				
Comparación A				
	SIMILITUD EWN	**% soporte Cortex**	**Soporte Cortex**	**Total parejas**
NUCLEO_SEXUALIDAD_A1	0.289	24.51	25	102
SATELITE_SEXUALIDAD_A1	0.235	13.13	13	99
NUCLEO_SUSHI_A1	0.136	15.62	15	96
NUCLEO_SUSHI_A2	0.1	40.91	18	44
NUCLEO_SUSHI_A3	0.163	15.76	32	203
NUCLEO_SUSHI_A4	0.185	14.29	7	49
NUCLEO_SUSHI_A5	0.393	22.22	8	36
NUCLEO_SUSHI_A6	0.073	17.14	6	35
NUCLEO_SUSHI_A7	0.214	10.00	2	20
SATELITE_SUSHI_A1	0.62	28.57	8	28
SATELITE_SUSHI_A2	0.084	16.67	3	18
SATELITE_SUSHI_A3	0.393	22.22	8	36
SATELITE_SUSHI_A4	0.297	16.67	5	30
SATELITE_SUSHI_A5	0.353	16.00	8	50
SATELITE_SUSHI_A6	0.533	13.33	2	15
SATELITE_SUSHI_A7	1	37.50	6	16
Promedio	**0.317**	**20.28**	**10.38**	**54.81**
Comparación B				
	SIMILITUD EWN	**% soporte Cortex**	**Soporte Cortex**	**Total parejas**
NUCLEO_SEXUALIDAD_B1	0.338	12.50	11	88
NUCLEO_SEXUALIDAD_B2	0.601	21.43	6	28
NUCLEO_SEXUALIDAD_B3	0.661	28.00	7	25
NUCLEO_SEXUALIDAD_B4	0.694	22.22	8	36
NUCLEO_SEXUALIDAD_B5	0.601	21.43	6	28
SATELITE_SEXUALIDAD_B1	0.694	22.22	8	36
SATELITE_SEXUALIDAD_B2	0.285	12.24	6	49
SATELITE_SEXUALIDAD_B3	0.714	33.33	2	6
SATELITE_SEXUALIDAD_B4	0.601	21.43	6	28
SATELITE_SEXUALIDAD_B5	0.885	40.00	8	20
NUCLEO_ASTRONOMIA_B1	0.524	16.67	7	42
NUCLEO_ASTRONOMIA_B2	0.259	9.23	18	195
SATELITE_ASTRONOMIA_B1	0.278	9.85	39	396
SATELITE_ASTRONOMIA_B2	0.292	16.57	28	169
NUCLEO_SUSHI_B1	0.244	23.93	28	117
NUCLEO_SUSHI_B2	0.229	12.50	3	24

NUCLEO_SUSHI_B3	0.244	23.93	28	117
NUCLEO_SUSHI_B4	0.667	16.67	2	12
NUCLEO_SUSHI_B5	0.286	16.33	8	49
NUCLEO_SUSHI_B6	0.332	21.37	50	234
SATELITE_SUSHI_B1	0.438	25.00	10	40
SATELITE_SUSHI_B2	0.328	18.75	3	16
SATELITE_SUSHI_B3	0.301	31.43	11	35
SATELITE_SUSHI_B4	0.667	16.67	1	6
SATELITE_SUSHI_B5	0.218	9.00	9	100
SATELITE_SUSHI_B6	0.148	11.37	29	255
SATELITE_SUSHI_B7	0.286	16.33	8	49
NUCLEO_SUSHI_B7	0.182	18.06	65	360
Promedio	**0.428**	**19.59**	**14.82**	**91.43**
Comparación C				
	SIMILITUD EWN	**% soporte Cortex**	**Soporte Cortex**	**Total parejas**
SATELITE_SEXUALIDAD_C1	0.067	15.87	10	63
SATELITE_SEXUALIDAD_C2	0.162	17.14	6	35
NUCLEO_SEXUALIDAD_C1	0.064	13.33	16	120
NUCLEO_SEXUALIDAD_C2	0.063	15.87	10	63
NUCLEO_ASTRONOMIA_C1	0	0.00	0	42
NUCLEO_ASTRONOMIA_C2	0.039	6.49	5	77
SATELITE_ASTRONOMIA_C1	0.025	5.68	5	88
SATELITE_ASTRONOMIA_C2	0.017	3.70	2	54
NUCLEO_SUSHI_C1	0.222	19.25	36	187
NUCLEO_SUSHI_C2	0.06	9.26	10	108
NUCLEO_SUSHI_C3	0.137	22.07	32	145
NUCLEO_SUSHI_C4	0.027	7.41	4	54
NUCLEO_SUSHI_C5	0.053	6.54	10	153
SATELITE_SUSHI_C1	0.145	12.50	10	80
SATELITE_SUSHI_C2	0.025	4.76	3	63
SATELITE_SUSHI_C3	0.074	16.67	6	36
SATELITE_SUSHI_C4	0.034	5.56	5	90
SATELITE_SUSHI_C5	0.054	12.50	22	176
Promedio	**0.070**	**10.81**	**10.67**	**90.78**